1977 - 1978

ANNUAL REPORT

RESOURCE-RELATED RESEARCH

COMPUTERS AND CHEMISTRY

Grant No. RR-00612

BIOTECHNOLOGY RESOURCES PROGRAM
OF THE
NATIONAL INSTITUTES OF HEALTH

February, 1978

COMPUTER SCIENCE DEPARTMENT
STANFORD UNIVERSITY

Resource Related Research - Computers and Chemistry
Stanford University
NIH/BRP Grant RR-00612

Carl Djerassi, Principal Investigator
(Social Security No. [redacted])

Research Highlights (1977-78)

1. Stereochemistry in Structure Elucidation.

The set of computer programs developed at Stanford as tools for molecular structure elucidation have been considerably enhanced by the addition of 3-dimensional structural information. The programs can now deal with some basic geometrical properties of molecules that are essential for understanding their biological significance. Research progress this year has resulted in extensions that allow computation of stereoisomers (alternative structures differing in 3 dimensions but having identical connections among atoms). Thus geometrical variations on structural hypotheses can be presented as well as topological variations.

2. Unified Package for Structure Elucidation.

Significant progress was made in unifying the computer programs for structure elucidation into a coherent package that is easily understood and used by chemists for complex biomolecular structure problems. Powerful tools are now well integrated for defining problem constraints, producing plausible solutions to structure problems, reducing the sets of alternative solutions with information about biosynthetic pathways, testing the alternatives, and suggesting new tests for further discrimination. New tools currently under development will be integrated into this same package.

Table of Contents

Resource Related Research - Computers and Chemistry

ANNUAL REPORT
August 1, 1977 - April 30, 1978

Stanford University
NIH/BRP Grant RR-00612

Carl Djerassi, Principal Investigator
(Social Security No. [redacted]

1 OVERVIEW OF RESEARCH ACTIVITIES

In this first year of a three year renewal, substantial progress was made on every major item in the renewal proposal. The most obvious facets of this interdisciplinary work on computers and chemistry are research, engineering and applications. On the research side, the computer programs have grown in both chemical and computer science sophistication. On the engineering side, the programs have been made faster and easier to use. On the applications side, the programs have been used by chemists working on biomedical problems at Stanford and elsewhere as aids in their own research (see [4]). In this report we stress progress along the dimension of research, but mention the other aspects in the discussions of research progress.

The report is organized by the following problem areas:

Structure Elucidation
Theory Formation
C^{13}-NMR Problems
Collaborative Research
Instrumentation

Unpublished work is discussed in some detail, while published work is summarized here. The project continues at a vigorous pace and remains an exciting research atmosphere because of the unique collection of researchers dedicated to the goal of producing intelligent computer aids for biomedical research.

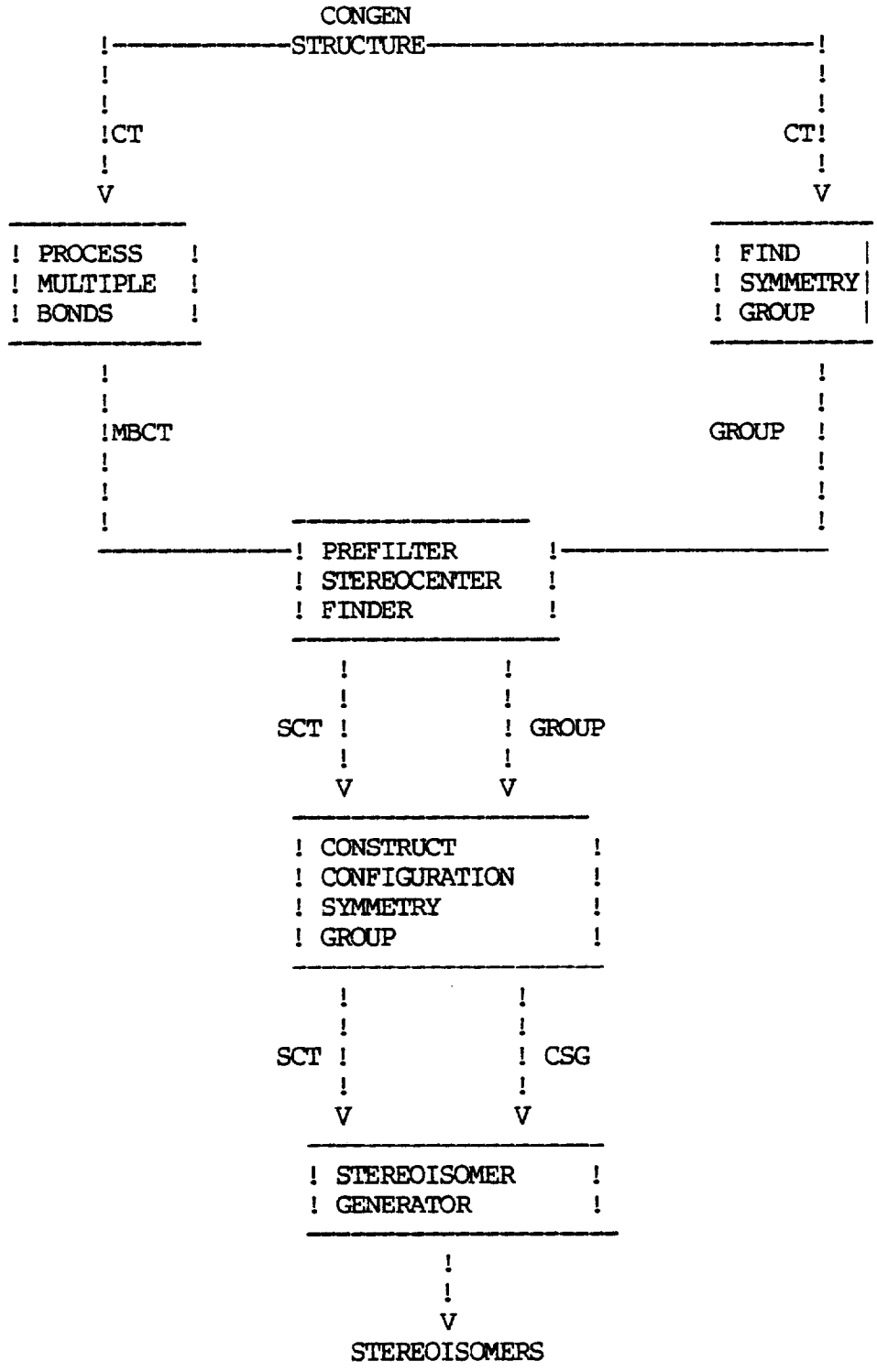
2 STRUCTURE ELUCIDATION PROGRAMS

2.1 Stereochemistry in CONGEN

The effort to give CONGEN the ability to recognize and use the stereochemical features of molecules in structure determination has continued for the past year. The proposed first stage in this effort was to write a program which was capable of recognizing the configurational stereochemical features of a molecule and generate all the possible stereoisomers based on these features. This program has been written and interfaced to an experimental version of CONGEN, and is described in detail below. The proposed second stage in this effort is to modify this program to permit generation of stereoisomers which satisfy certain constraints, much as the existing CONGEN program constrains the generation of topological isomers. This ongoing effort is discussed in the section on future plans.

Each module of this program, written in SAIL, is described in detail below. In summary, the program takes a structure defined in CONGEN and extracts the Connection Table (CT) from it. The symmetry group of this structure is found based on this connection table. The CT is then searched for features corresponding to multiple bond stereo features (double bonds, allenes, etc.) and the CT is modified to the Multiple Bond Connection Table (MBCT). Making use of the symmetry group, the MBCT is then searched for stereocenters (asymmetrically substituted carbon atoms, etc.) to yield the Stereochemical Connection Table (SCT). Using the SCT, the symmetry group is modified to recognize the effect of the symmetry operations on these stereocenters. The resulting group is the Configuration Symmetry Group (CSG). The SCT and the CSG are then used together to generate the possible stereoisomers for the input structure. These are output with other information in the manner described below.

Stereoisomer Generator Program


CONGEN
STRUCTURE
CT
CT
PROCESS MULTIPLE BONDS
FIND SYMMETRY GROUP
MBCT
GROUP
PREFILTER STEREOCENTER FINDER
SCT
GROUP
CONSTRUCT CONFIGURATION SYMMETRY GROUP
SCT
CSG
STEREOISOMER GENERATOR
STEREOISOMERS

2.1.1 Process Multiple Bonds

This module takes the CT and converts it into a Connection Matrix (CM) for use here and in the group finder described below. The CM is searched for all double and triple bonds. The atoms involved in triple bonds are flagged as stereochemically uninteresting. Double bonds and cumulenes with CH2 ends are similarly flagged. All remaining doubly-bonded atoms are potential stereocenters at this stage. These are processed by attaching a fictional bivalent node to each edge of the double bond, thus giving the multiply-bonded atom four distinct neighbors which aids in configuration assignment and in representation of the permutation group. These fictional nodes are given numbers higher than those already used in the structure and the corresponding rows are added to the connection table, to yield the Multiple Bond Connection Table (MBCT). (See examples.)

2.1.2 Find Symmetry Group

This module finds the node symmetry group of the input CT and was constructed largely of existing code from other parts of CONGEN, thereby saving the time and effort of developing another large program. This segment can be used independently from the rest of the program, a useful feature since previous group finders were written for very specific purposes. The symmetry group is constructed in two parts. The first is the node symmetry group of the input CT. The second is the symmetry group associated with the fictional nodes which were added to the MBCT described above. These two groups combine as a semidirect product. However, the utilization is such that the product group never needs to be explicitly constructed. This means the group can be stored in two arrays of size nXp and fXq where n is the number of original nodes, p is the order of the node symmetry group, f is the number of fictional nodes and q is the order of their symmetry group. If the entire group were constructed, the storage array would be of size nfXpq. Since the symmetry group can be by far the largest data structure in the program, the saving of space by this technique is crucial.

2.1.3 Prefilter

This module is all new code which recognizes all the stereochemically interesting features of the input structure based on the configuration of tetravalent atoms. The program works backwards by rejecting all those atoms which can never exhibit configurational stereochemistry. The MBCT is scanned first to eliminate all methyls and methylenes from further consideration as stereocenters. These atoms are flagged as nonstereocenters. Following this, all atoms with symmetrically related substituents are found using the node symmetry group described above. A crucial feature here is that the parity (odd

or even nature) of the permutations must be recognized and only odd permutations are considered. It is this property which leads to many of the seemingly pathological cases that confound many attempts at rigorous description of stereochemistry. Having done this each potential stereocenter with symmetrically related substituents is checked to see if those substituents themselves contain potential stereocenters. If they do not, then the node to which they are attached can never exhibit configurational stereochemistry and is flagged as such. Thus a carbon atom with two methyl substituents would be found not to possess stereochemistry in this way. The procedure of checking potential stereocenters is done iteratively, as long as new nonstereocenters are found. Since multiply-bonded atoms have already been processed to look like tetravalent saturated atoms, they are treated similarly here. The output of this module is the Stereochemical Connection Table (SCT) which includes only those atoms which are capable of exhibiting configurational stereochemistry. Atoms which were rejected as stereocenters by this module are retained for use in reducing the size of the relevant symmetry group as described in the next section. Since the number of potential stereoisomers increases as 2^m where m is the number of potential stereocenters, reducing the size of m to the minimum necessary is a substantial efficiency both in time and storage. (see examples)

2.1.4 Configurational Symmetry Group

The purpose of this module is to determine the effect of the permutations in the symmetry group on the potential stereocenters. This representation of the symmetry group is necessary for the generator to work properly. The basic part of this module is largely unchanged from last year's version as described in the previous annual report. Two modifications have been made since then. The first is that the symmetry group is processed here as elsewhere in the program as two separate pieces for the reasons described above. Second, it was found that a substantial saving could be made by reducing the size of the symmetry group to that subgroup (technically a homomorphic image) which is concerned only with the potential stereocenters. This is done by eliminating those permutations which only effect parts of the molecule which do not exhibit any configurational stereochemistry. Since these parts of the molecule were themselves found earlier by just these permutations, it is a relatively easy matter to discard them afterwards. The resulting symmetry group is reduced by(at least) a factor proportional to 2^r where r is the number of "rejected stereocenters". This leads to a significant savings in time since the symmetry group must be scanned through several times when stereoisomers are generated.

2.1.5 Generator

This module takes the SCT and CSG and generates all the possible stereoisomers. The basic workings of this program are as described in the previous annual report. Modifications were necessary to accommodate the two part symmetry group as described above. Two new features have also been added here. First, the program is capable of detecting enantiomeric pairs of stereoisomers based on the configuration of the stereocenters. This does not include cases where enantiomerism results from conformational or other structural features. Second, the program is capable of computing the symmetry group of each stereoisomer. In general this will be a much smaller group than the CSG for each individual stereoisomer. These two features were added in anticipation of their need later on when capabilities for constrained stereoisomer generation become available. Interpretation of spectral properties such as proton and carbon nmr generally require knowledge of the symmetry group of the stereoisomer being examined. At this stage the outputted stereoisomer is in a canonical form based on the input numbering of the original CT. Because of the very compact representation possible for stereoisomers discussed in last year's annual report, this canonical form is simply an integer from 0 to 2^n where n is the number of stereocenters. Some future plans for the more transparent output required are discussed in the section on future plans. (See example.)

2.1.6 Examples

Several examples are provided here to demonstrate some of the capabilities of the program.

Example 1. The first is 3-6-dimethyl-4-octene, a simple hydrocarbon which exhibits double bond and configuration stereochemistry and has a reduced number of stereoisomers due to symmetry.

3-6-dimethyl-4-octene

```
1 0 0 0 2
2 0 0 1 3
3 0 5 4 2
4 0 3 0 0
5 0 3 11 12
6 11 12 7 0
7 6 8 9 0
8 0 0 0 7
9 0 0 7 10
10 0 0 0 9
11 5 6 0 0
12 5 6 0 0

 THE SCT:

3 0 5 4 2
7 6 8 9 0
5 0 3 11 12
6 11 12 7 0

STEREOCOUNT= 6

 THERE ARE 6 STEREOISOMERS
0 1 1 -1
1 0 1 1
2 0 1 1
4 1 1 -1
5 0 1 1
6 0 1 1
```

Five separate output results are given for this example:

1) The first twelve rows are the Multiple Bond Connection Table (MBCT). The first number is the atom number and the following four are the atoms to which it connects. (0 is hydrogen) Rows 11 and 12 are correspond to the fictional nodes which label the edges of the double bond.

2) Next is shown the Stereochemical Connection Table (SCT). The program has found the two asymmetrically substituted carbons (3 and 7) and the double bond (5 and 6).

3) A counter (discussed below) has determined that there are 6 distinct stereoisomers. This is the STEREOCOUNT.

4) The generator has likewise determined that there are 6 stereoisomers.

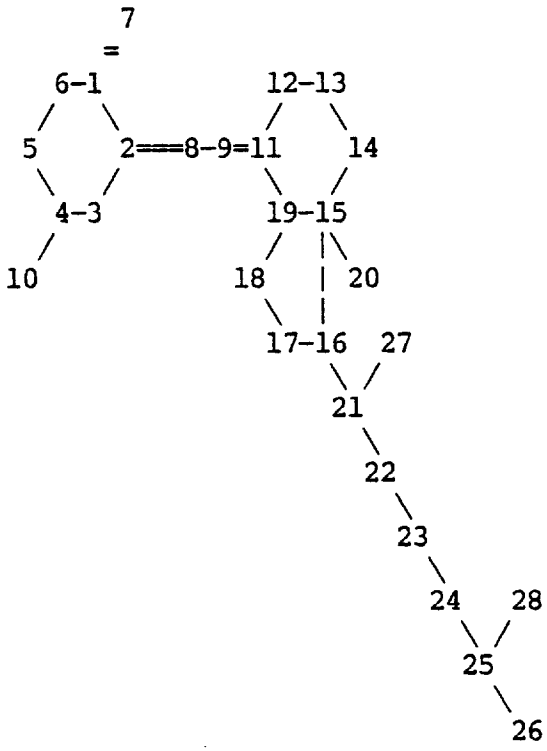
5) The stereoisomers are listed. The first number on each row is the canonical label for each. The correspondence is:

0 R-S-trans
1 S-S-trans
2 R-R-trans
4 R-S-cis
5 S-S-cis
6 R-R-cis

The second number on each row tells whether this particular stereoisomer is achiral (1) or has an enantiomer (0). Enantiomeric pairs are listed on consecutive rows. The final two numbers on each row indicate the symmetry group of each stereoisomer. Those with 1 1 have rotational symmetry and those with 1 -1 have a plane of symmetry.

Example 2. The second example is Vitamin D3 and is included here to illustrate the capabilities of the program in finding stereocenters.

Vitamin D3



Atom number 10 is Oxygen, the rest are Carbon.

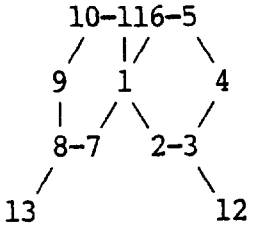
THE SCT:

```
4 10 3 5 0
15 20 19 14 16
16 21 15 17 0
19 15 11 18 0
21 27 16 22 0
2 29 30 1 3
8 29 30 9 0
9 31 32 8 0
11 31 19 32 12
25 28 24 26 0
```

STEREOCOUNT= 128

THERE ARE 128 STEREOISOMERS

For this example only the SCT and number of stereoisomers are shown. The first 5 rows correspond to the 5 asymmetrically substituted carbons. The next four rows correspond to the 4 doubly-bonded atoms which can exist in distinct cis and trans forms. The final row corresponds to the gem-dimethyl substituted carbon on the side chain. This is retained for the reasons discussed above. Both the counter and the generator have established that there are 128 stereoisomers (the theoretical maximum). Example 3. The disubstituted spiro-undecane shown below has only one element of symmetry, the "rotation" axis through carbon 1. This is an even permutation so that carbon 1 remains a stereocenter. NST is the number of stereocenters, NDBAT is the number of doubly-bonded atoms and NRJ is the number of stereocenters rejected by the prefilter.



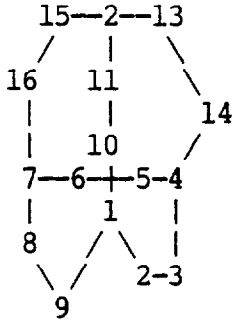
THE SYMMETRY GROUP HAS ORDER P= 2
NST= 3 NDBAT= 0 NRJ= 0

STEREOCOUNT= 6

THERE ARE 6 STEREOISOMERS

```
0 0
7 0
1 0
6 0
2 0
3 0
```

Example 4. The hydrocarbon shown below is the higher homolog of adamantane. The conformational process of turning the structure "inside-out" interconverts the structure with all the hydrogens pointing inside the cage with the structure with all the hydrogens pointing out. The same process interconverts the 3 out 1 in structure with the 1 out 3 in.

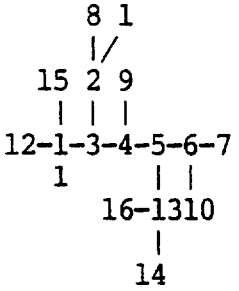


```
THE SYMMETRY GROUP HAS ORDER P= 24
 NST= 4 NDBAT= 0 NRJ= 0

STEREOCOUNT= 3

THERE ARE 3 STEREOISOMERS
 0 1
 1 1
 2 1
```

Example 5. The substituted heptane shown below has two extensively branched symmetrically related substituents at the central carbon. The program detects that this structure can have only 1 stereoisomer and prints this out rather than going through the counting and generating procedures.



```
THE SYMMETRY GROUP HAS ORDER P= 128
 NST= 0 NDBAT= 0 NRJ= 7 THERE IS 1 STEREOISOMER
```

2.1.7 Counter

Another new feature of the program is a procedure which counts the number of stereoisomers for a structure without generating them by using the CSG and the appropriate

combinatorial theorem. This represents the first solution to the problem which dates back to the 1870's. Since the counter works much faster than the generator, this is a very useful feature as the number of stereoisomers can be obtained quickly if only this is needed. This differs from the structure generator where a faster counter was not possible. In addition, having the counter and the generator working independently allows a mutual checking for bugs during development of the program since the two results must be the same for any test case.

2.1.8 Interface to CONGEN

The current interfaced version of the stereogenerator with CONGEN is intended primarily for testing purposes and does not represent the final version. The stereogenerator runs as a separate SAIL fork which is started only when the STEREO command is issued. The desired structure is constructed as a pattern in EDITSTRUC. The CONGEN command: STEREO (name) starts the fork and the stereogenerator. The program asks for an output file and then returns a brief summary of the results to the terminal and a more complete set of results is written on the file. On termination of the generator, control returns to CONGEN.

2.1.9 Future Plans

The following features (at least) will be added to the existing program:

1) Designations of stereocenters as either R or S based on constitutional priorities only. This will be for aid in interpretation only as these designations are not useful internally to the program.

2) Recognition of cis and trans double bonds for the same reason.

3) Stereoisomer output which is interpretable and compatible with character terminal output. This will most likely be done in conjunction with the existing drawing program. The compatibility with character based terminals is a strength of CONGEN at present.

4) Versatility in the handling of the stereochemistry of atoms other than carbon. In particular there should be a choice as to whether a nitrogen atom is thought to be able to invert freely.

The second stage of the development in this effort is to give CONGEN the ability to constrain stereoisomer generation. The algorithm of the generator was designed so that a number of useful constraints, particularly concerning relative

stereochemistry between stereocenters can be applied prospectively. That is, the undesired stereoisomers would not be generated. Other constraints, such as those which involve the symmetry of the stereoisomers can be applied during the generation. Finally, there will certainly be some constraints which have to be applied after generation.

2.2 Constraints Interpretation

The area of automatic interpretation of constraints in CONGEN structure elucidation problems is interesting and important for two reasons: 1) we want to free the chemist as much as possible from having to understand CONGEN's method of building structures; and 2) problems can be solved much more efficiently if CONGEN can perform some preliminary examination of them and find an alternative, efficient way to solve the problem. Our first efforts in this direction have resulted in what we call the "GOODLIST interpreter", which employs the method of constructive substructure search as described in the following sections. The GOODLIST interpreter is designed to make more efficient use of information about required (GOODLIST items plus Superatoms) structural features of an unknown molecule.

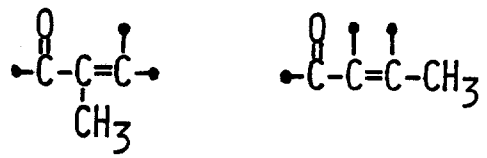
2.2.1 Abstract of Method

We present a solution to the problem of constructing all structural isomers of a given empirical formula given also a set of required partial structures which overlap, i.e., share atoms in common, to an unknown extent. Our method takes a collection of non-overlapping partial structures (in the limit, all atoms in the empirical formula) and, using a technique we term "constructive substructure search," determines the set of subproblems which incorporate all given partial structures, including all possible overlaps, required to be present in each isomer. Each subproblem is solved in turn by CONGEN to yield finally the complete set of isomers, e.g., structural candidates for an unknown compound. Our method allows facile solution of certain structural problems which are beyond the scope of other computer-based methods.

2.2.2 Introduction to Method

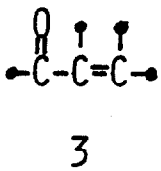
It is characteristic of structure elucidation based on data from physical and chemical methods that much structural information is redundant. Physical methods, for example, are frequently complementary. One technique provides structural information which can be used to elaborate information gathered by another. The collection of partial structures present in an unknown derived by such methods frequently contain atoms or

groups of atoms shared among two or more partial structures. Chemists must take this into account when considering how the partial structures might fit together to yield the structure of an unknown compound. As a simple example, the carbon-carbon double bond of an inferred vinyl methyl functionality may or may not be the same as the double bond of an inferred , -unsaturated ketone. As long as the empirical formula admits of two (or more) double bonds and in the absence of additional information, both possibilities must be considered. Therefore, the chemist will consider 1, 2 and 3,4 as tentative building blocks for further elaboration of the example structure.

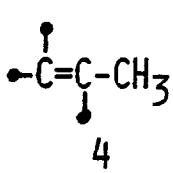


1

2



3

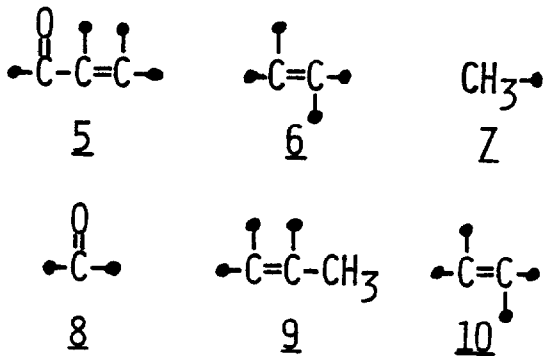


4

Although computer programs, including CONGEN, now exist to assist chemists in constructing structural isomers based on information about partial structures, the programs have one serious limitation in common. Each program must use as building blocks non-overlapping structural fragments. This limitation leads to at least two important problems; 1) The chemist using such a program must select non-overlapping partial structures; otherwise an incomplete set of structures will result. This manual procedure is time-consuming, unnatural and prone to error; and 2) as a consequence of (1), problems are solved less efficiently by the program because the detailed environment of fewer atoms is specified to ensure the absence of overlaps. Thus, undesired structures are built only to be discarded upon later evaluation. We feel that a solution to the first problem is extremely important. Our experience is that there are already sufficient barriers to use of computers as assistants in problem-solving. We feel strongly that allowing a chemist to input structural information freely without regard to overlapping partial structures would reduce that barrier. The importance of the second problem is that certain structural problems become difficult or impossible to solve with current programs (that is,

impossible in the sense that resources of computation, time and money are finite).

For the example cited above, current programs would be forced to consider for completeness a starting point of either 5,6,7 or 8,9,10.



Assuming that the problem involves other partial structures or atoms, either starting point results in construction of structures including 1, 2 and 3,4 together will many other structures which do not obey the constraints on the problem. Application of constraints in CONGEN is automatic, but the retrospective testing of every structure for desired structural features which could not be used to begin with is very inefficient.

We sought, therefore, a method which would emulate the manual approach to the problem of determining structural candidates based on overlapping partial structures. Stated in the simplest terms, the method should translate the constraints on desired structural features, or GOODLIST constraints, into new sets of partial structures which incorporate the features at the beginning of the structure generation procedure. Such a method would translate automatically the constraints in the problem mentioned above to yield three new problems represented by 1, 2 and 3,4. Subsequent sections describe a method which performs this translation. We illustrate the method with examples drawn from our own work, some of which could not be solved in reasonable time using existing programs.

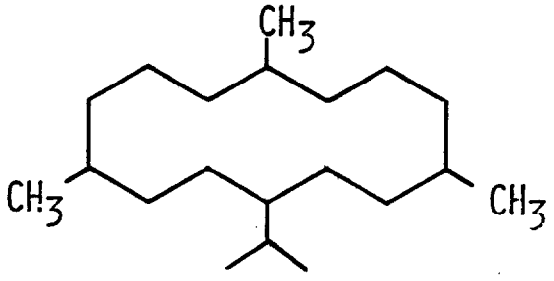
2.2.3 METHOD

There are usually many constraints on a structural problem brought to CONGEN, including those implied by other constraints. Manual approaches to structure elucidation involve recognition of implied constraints and resolution of overlapping partial structures (mentioned above) as structural candidates are constructed. The translation of constraints to discern their implications and elaboration of those implications into more efficient statements of a problem involves complex reasoning about chemical structures. This reasoning is susceptible to analysis and encoding in a computer program.

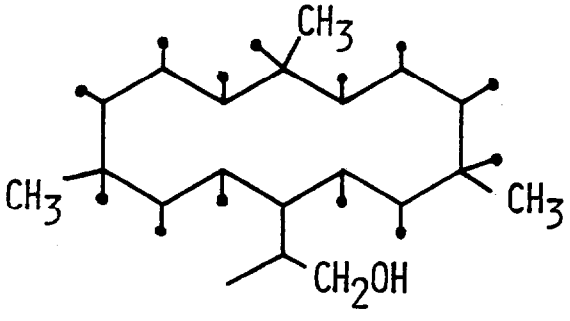
Our initial experiment in constraints interpretation involved determination of the implications of designated numbers of hydrogens associated with particular atoms. Translation of this information reduces many problems to triviality, for example, "construct all isomers of $C_{20}H_{44}N_2$ which possess no methyl groups". We describe below the next step in our efforts, a method for translation of desired, or GOODLIST, structural features.

Our method is based strongly on our observations of how chemists actually solve the problem of using overlapping partial structures. We introduce the method with an example which in fact provided the basis for the first programming efforts.

The structural problem involved an unknown compound of empirical formula $C_{20}H_{34}O_1$. The compound was isolated together with other cembranolides, therefore the assumption was that the unknown possessed the unrearranged cembrane skeleton (11).



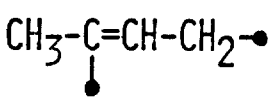
11



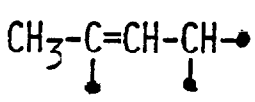
12

These data indicate that the structure is based on the

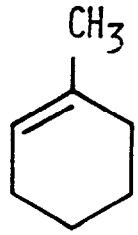
skeleton 12 together with allocation of three new bonds in such a way as to yield the desired partial structures 13-15. (Bonds with an unspecified terminus, or "free valences" in 12 may be to any atom including hydrogen, while in 13-14 the indicated free valences are specified to be to non-hydrogen atoms.)



13



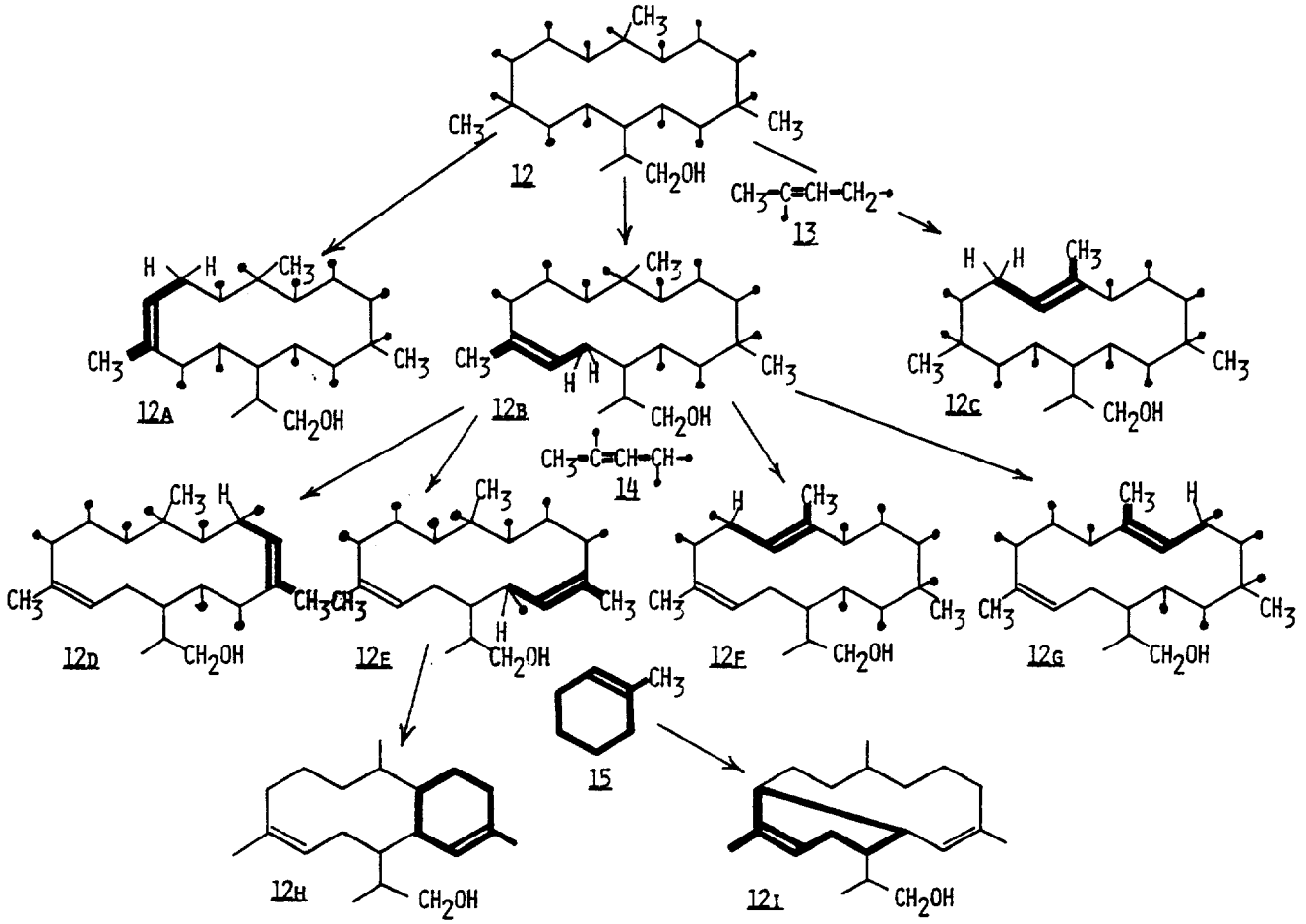
14



15

In this problem, the skeleton, 12, possesses all non-hydrogen atoms of the empirical formula. Thus, the substructures 13-15 overlap completely with 12 (and partly with each other). A conventional approach to this problem would allocate three new bonds to 12 in all possible ways and test each result against the GOODLIST constraints 13-15. There are many thousands of possible allocations and the computational task of building and testing each one was so time consuming it was terminated. The chemist then retired to his desk and, using pencil and paper, in a short time determined the seven possible structures obeying the constraints.

It is clear conceptually how such problems are solved. It is obvious, considering the topological symmetry of 12, that there are only three places in 12 where 13, for example, might fit, or match. The three matchings 12a-12c are shown below. Each matching consumes two free valences to form the new double bond and effectively places a hydrogen on the terminal atom of the substructure yielding the required $-CH_2-$ group. For each matching of 13, there are several ways to fit in the next GOODLIST substructure, 14. There are four ways to perform this matching for 12b, resulting in 12d-12g, below. Again, a pair of free valences is consumed to construct the new double bond. In this case, however, the substructure 14 terminates in a methine group, effectively leaving a bonding site open (see 12d-12g) which must be used in forming a new bond in a subsequent step. Incorporation of the final GOODLIST constraints, 15, proceeds by creation of a new bond (with the methine, above, as one terminus) to yield a six-membered ring possessing a double bond. Certain structures, e.g., 12f, yield no results because a bond cannot be formed which meets the requirements of 15, while 12g yields two results 12h-12i, as shown below.



In this example, some matchings result in construction of new bonds to form the extra double bonds and ring of the unknown. In the general case, the procedure is constructive in that bonds are formed to new atoms or substructures to obtain partial structures which are required. Using the method described below in conjunction with CONGEN, we can determine automatically and quickly the seven solutions.

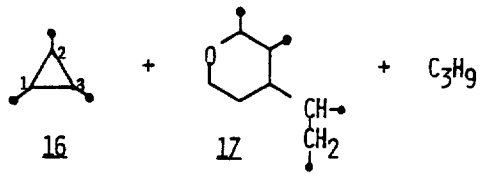
2.2.4 GOODLIST Constraint Interpretation Search

Our method emulates the manual method by searching for ways to map possibly overlapping GOODLIST substructures into the partial structures and/or atoms in the initial problem formulation. The method, illustrated schematically below, includes the following steps.

2.2.4.1 Formulation of the Initial CONGEN Problem

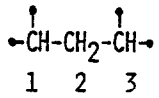
The initial structural problem is defined to be a set of non-overlapping partial structures, or "Superatoms," plus the remaining atoms in an empirical formula (below). Thus, specifications of the initial problem can proceed just as with current use of the program. However, a wide variety of initial specifications is possible, from initial problems where all atoms are part of a superatom (e.g., 12, above) to the limit of simply the empirical formula (where all atoms are of course non-overlapping). For example, the problem of the cembrenolide outlined above is solved with little difference in efficiency beginning with the empirical formula and utilizing 13-15 as GOODLIST constraints. In the example below, assume that partial structures 16 and 17 are known to be non-overlapping superatoms, leaving C_3H_9 remaining from an empirical formula $C_{13}H_{22}O_1$.

INITIAL CONGEN PROBLEM



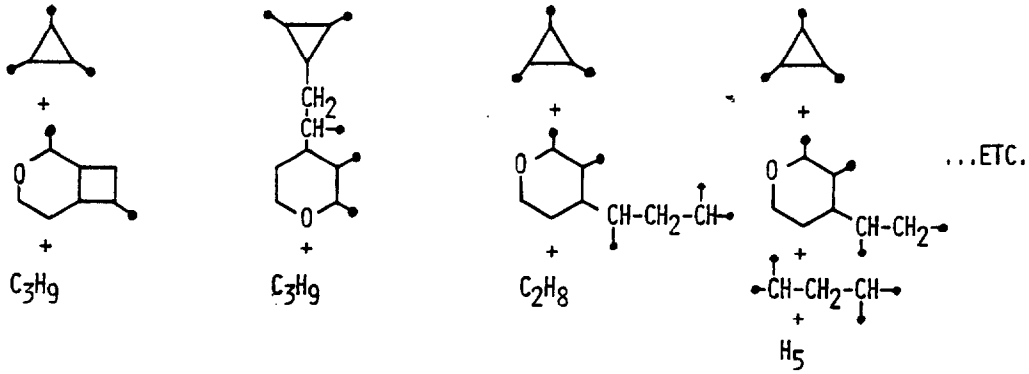
16 17

GOODLIST CONSTRAINT



18

NEW CONGEN PROBLEMS



A B C D

2.2.4.2 Constructive Substructure Search

Assume that substructure 18 is known to be present in a molecule of unknown structure with no additional information on possible overlaps with 16 and 17. The method begins by finding

all ways in which the GOODLIST substructure (18) can be constructed using Superatoms and atoms in the initial problem.

There may be several ways to incorporate a given GOODLIST constraint in a CONGEN problem. The substructure may be incorporated by forming bonds within a substructure (yielding A), forming new bonds between (or among) substructures (yielding B), forming bonds between substructure(s) and remaining atoms (yielding C) or construction of the substructure wholly from remaining atoms (yielding D).

The result of constructive incorporation of each GOODLIST substructure is a set of new CONGEN problems. Our stepwise procedure continues by incorporating the next GOODLIST item in a depth-first generation scheme. For example, considering the cembrenolide, above, one of the three new problems after incorporation of 13 is chosen for the next step, incorporation of 14. One of the resulting problems is chosen for incorporation of 15. The procedure continues until all GOODLIST items have been incorporated or until the next GOODLIST item cannot be built from superatoms and atoms in the current problem. In the latter case, the program backtracks one step and tries the next problem at the previous level.

2.2.4.3 Obtaining Final Structures

The results of the constructive procedure may be complete structures, for example, 12h and 12i. Usually, however, the result is a set of incomplete problems. Each problem includes superatoms and remaining atoms which are guaranteed to be non-overlapping and which contain all desired structural features. The standard CONGEN procedure for structure generation can then be invoked. However, the task of testing for substructure and ring constraints is simplified in that GOODLIST constraints are already incorporated.

2.2.5 Limitations

There are some limitations to the procedure which decrease its efficiency compared to what might be possible with further work. One limitation is the problem of duplication inherent in the procedure. Although many steps are taken to perceive and utilize topological symmetry in the constructive substructure search, there remains the possibility of constructing duplicate CONGEN problems whenever the constructive procedure creates symmetries which were not present originally. Therefore, we convert each CONGEN problem to a canonical form and compare problems to eliminate duplicates. Another potential source of duplication is construction of duplicate (isomorphic) final structures from different CONGEN problems. Again, canonicalization serves to prevent presentation of duplicate structures to the chemist.

A second limitation is related to the absence of a mechanism for preventing the association of atoms in a GOODLIST substructure with atoms in a CONGEN problem. It may be known that a GOODLIST substructure does not share atoms (i.e., overlap) with one or more superatoms (i.e., some spectroscopic evidence is available to distinguish them). However, there is no mechanism for preventing association of atoms in a superatom with atoms in a GOODLIST item. Some undesired structures result which must be removed by subsequent tests.

2.2.6 Future Directions

The program described in this section will be incorporated in the existing CONGEN program in such a way that it will be invisible to the chemist using the program. Initially, the GOODLIST substructures specified as constraints will be incorporated automatically at the beginning of the problem as described above. Within a short time, the method of specification of a problem will be changed to include only the empirical formula together with inferred partial structures without regard to overlaps, leaving to the program the task of determining those overlaps and specifying the set of problems to solve.

Automatic interpretation of GOODLIST constraints is only the first phase of our efforts. Incorporation of BADLIST (undesired structural features) substructures in the procedure is a necessary next step. Subsequently we will attack the problem of discerning constraints which are implied by the input data, including detection of unclear or ambiguous statements about a structure. The constraints interpreter should be capable of a dialog with the chemist using CONGEN to clarify such points prior to structure generation.

2.3 Experiment Planning Program

Now that Congen gives us the capability of constructing all plausible candidates under an initial set of constraints, the next problem is to provide the chemist with some assistance in rejecting incorrect candidates and focussing on the correct structure. This process must involve the examination of the candidates to determine their common and unique features, and the designing of experiments to differentiate among them.

The initial work on this problem has begun by providing a new function, the EXAMINE function, which gives a chemist the ability to survey sets of structures for particular combinations of substructures, ring-systems etc. This function has now been incorporated into the CONGEN program; details and examples are given later.

More elaborate functions for automatically identifying discriminating features in sets of structures are being developed. Currently, these experimental routines (contained within the "PLAN" program) can be used to analyze functionality, or to identify differences in the ways that superatoms have been imbedded in structures. These routines will shortly be capable of exploiting a simplified library of chemical/spectral tests for particular substructural features; this will allow the program to identify possible discriminating experiments. The current capabilities of these functions are described in subsequent sections.

2.3.1 EXAMINE

The EXAMINE function allows for the identification and selection of structures characterized by particular combinations of substructures, ring-systems and Isoprene-patterns. Further, if relative merits can be associated with the substructural features, then these merit values can be used to rank the structures. In addition to providing information on the frequency of different structural features, the EXAMINE function allows structures with unacceptable combinations of features to be pruned away.

EXAMINE thus extends both the earlier SURVEY function (which EXAMINE has now totally subsumed) and the PRUNE function in CONGEN. (PRUNE remains in CONGEN because of its greater efficiency in simply rejecting undesired structures.) EXAMINE allows structures to be segregated on the basis of combinations of (desired or undesired) structural features. For example, EXAMINE can be used to segregate structures which possess feature A or B, or generally, any arbitrary Boolean expression of relationships among structural features.

The EXAMINE function involves the following steps:

1) the definition of relevant substructural features.

2) [EXAMINE matches the features to the structures produced by an earlier GENERATE or IMBED step, and summarizes their frequency.]

3) [if some form of merit rating is being used, then details of the ranking process are provided.]

4) then, in "EXAMINE sub-command" mode, subsets of structures possessing different combinations of features may be selected. Features may be combined using standard AND/OR/XOR/NOT operators. These subset selection procedures are basically non-destructive; however, it is possible to use them to prune the structure list.

5) if examination of the structures has suggested additional selection features, then the entire process may be repeated (information on the current selection features being preserved to allow new selection criteria to be combined with those already in existence). Previously defined libraries of selection features can be used, either alone or as a supplement to selection features specified for a particular problem. It is also possible to save the current set of selection criteria for future use.

2.3.1.1 Example - Unknown Metabolite from Human Urine

Use EXAMINE to determine which members of a set of candidate structures possess naturally occurring, alpha-amino acid part structures. The compound for which CONGEN provided structural candidates was an unknown component of human urine. The empirical formula was $C_{15}H_{19}NO_5$. There were 78 structural candidates based on this empirical formula and chemical constraints. Ten of the 78 formally possess an alpha-amino acid substructure (-NHCHCOO-). Examination of these structures proceeded as follows (note that the examination would yield the same results if the entire 78 were examined).

```
 EXAMINE
Do you require simply to prune your structure list?:
Do you want to rank your structures?(Y for Yes, ? for
   explanation):
Do you want to use a library?Y
FILE NAME:AMINOACID.LIBRARY;8 [Old version]
READING <SMITH>AMINOACID.LIBRARY;8
Do you want all substructures in the file?:Y

      (file read OK)
Do you want to enter new selection features?:
ALA-1-? Substructure ALA min/max (1 . ANY) present in 1
   structures.
GLY-1-? Substructure GLY min/max (1 . ANY) present in 0
   structures.
VAL-1-? Substructure VAL min/max (1 . ANY) present in 0
   structures.
LEU-1-? Substructure LEU min/max (1 . ANY) present in 0
   structures.
ILEU-1-? Substructure ILEU min/max (1 . ANY) present in 0
   structures.
THRE-1-? Substructure THRE min/max (1 . ANY) present in 0
   structures.
PHE-1-? Substructure PHE min/max (1 . ANY) present in 2
   structures.
TYR-1-? Substructure TYR min/max (1 . ANY) present in 0
   structures.
PRO-1-? Substructure PRO min/max (1 . ANY) present in 0
   structures.
OH-PRO-1-? Substructure OH-PRO min/max (1 . ANY) present in 0
   structures.
```

```
ASP-1-? Substructure ASP min/max (1 . ANY) present in 1
   structures.
GLU-1-? Substructure GLU min/max (1 . ANY) present in 1
   structures.
BETA-ALA-1-? Substructure BETA-ALA min/max (1 . ANY) present
   in 0 structures.
SER-1-? Substructure SER min/max (1 . ANY) present in 0
   structures.
```

[note that only four of the amino acids have their part structures (-NHCHR-COO-) represented in the set of candidates, alanine (ALA), phenylalanine (PHE), glutamine (GLU) and asparagine (ASP)]

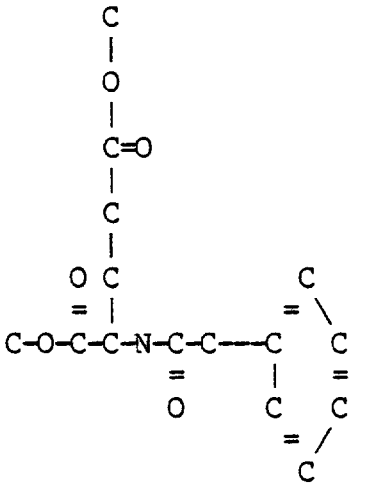
```
Enter commands for selecting subsets of structures with
   particular features.
Do you want help?:
10 STRUCTURES
->SELECT
>(ALA-1-? OR PHE-1-? OR ASP-1-? OR GLU-1-?)

    ...
5 STRUCTURES WITH ((ALA-1-? OR PHE-1-? OR ASP-1-? OR GLU-1-?))
```

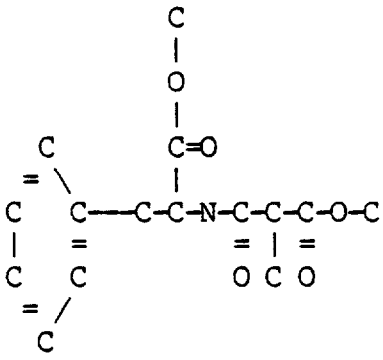
[Only five of the ten (or 78) have any one of the four amino acid substructures. They are drawn below. The first structure drawn is the 77th of the 78 original candidates. The second number refers to its rank based on a comparison of the mass spectrum predicted for this compound against that observed for the unknown. This compound was among the three top-ranked structures (MSRANK) in the original set of 78. It is clearly ranked higher than the other four candidates under the (biochemical) constraint that the compound contain the substructure of a naturally occurring amino acid. Subsequent synthesis and comparison of GC and MS confirmed the identity of the unknown as phenylacetylglutamic acid dimethyl ester.]

->DRAW

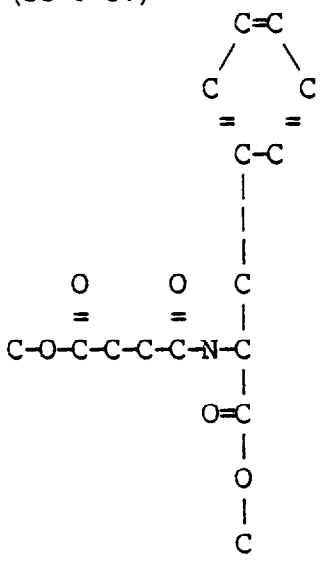
(77 . 84)



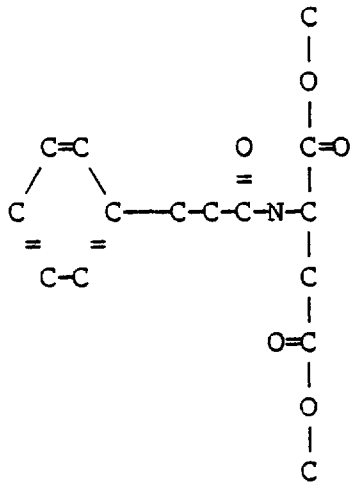
(57 . 57)



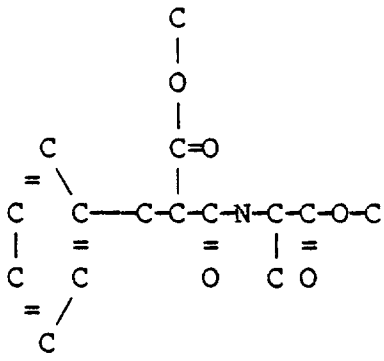
(55 . 57)



(51 . 57)



(29 . 57)



->DONE

2.3.2 PLAN

As mentioned previously, the PLAN program represents our initial efforts toward assembling the heart of an experiment planning program. The goal of PLAN is to identify all structural features which distinguish among structural candidates for an unknown. In the next year we will develop the program which will use this information to suggest experiments. The EXAMINE function, described above, can only look for structural features explicitly supplied by the chemist. Although a summary of such features is quite useful, EXAMINE is insufficient to solve the more general problem of identifying distinguishing substructures.

PLAN in its current form provides the following capabilities:

1) Using a starting substructure supplied by the chemist (for example, one of the superatoms used to construct structural candidates), PLAN can search the local environment of the substructure for distinguishing features, continuing the search until discriminatory characteristics are found.

2) PLAN checks (if requested) for simple differences in the distribution of carbon and hydrogen atoms which could be detected by $^{13}C_{MR}$ or $^{1}H_{MR}$.

3) PLAN can begin at existing functional groups and examine larger substructures by expanding the local environment (as in (1), above) until distinguishing features are found. The example below represents PLAN operated in this mode.

4) PLAN, if requested, performs the operations specified in (3) beginning with double bond systems in the candidates.

2.3.2.1 Example

In the following example, 88 structural candidates for the compound palustrol[8], based on spectroscopic information, were processed by PLAN. The following is a recording of that terminal session. Bracketed comments ([ ]) are inserted to explain the flow of the program.

```
@congen                                              [begin CONGEN]
(<SMITH>CONGEN.;22 . <LISP>CARHART.SAV;70702)
:OK
(LISP)
DO YOU WANT TO SPECIFY AN EMPIRICAL FORMULA?(Y FOR YES):
 RE                                    [RESTORE file of structures]
INPUT FILE:PAL.REACT [Old version]
READING <SMITH>PAL.REACT;2
 THIS IS A FILE WRITTEN BY CONGEN
(COMPOSITION RESTORED)
(EMPIRICAL FORMULA RESTORED)
(AROMATICS RESTORED)
(CONSTRAINTS RESTORED)
USERATOMS  HEP  A1  B1  CH3  CH2  CH  ETH  MET  C  N  O
ALL RESTORED
(88 STRUCTURES)                                      [88 candidates]
 LISP

(LISP):
:(PLAN)                                              [Begin PLAN]

Do you want to specify starting superatoms?     [No starting
                                                 point specified]
Do you want  the program to check  for simple differences  in the
off-resonance decoupled 13c spectrum?Y
```

These structures show no simple differences in their carbon distributions.

Do you want the program to check for simple differences in proton distributions? Y

These structures show no simple differences in their hydrogen distributions.

Do you want the program to check functional groups?Y (See mode (3), above)

Only one substructural class was generated

All compounds have this feature:

```
OH-C
                                    [All compounds possess a
                                    tertiary hydroxyl group,
                                    so PLAN continues]
OH-C
```

present in 88 structures

Only one substructural class was generated

All compounds have this feature:

```
  C
  |
C-C-O                [All compounds have three carbon atoms
  |                  bonded to the tertiary OH, but the
  C                  hydrogen distributions on those carbons
                     differ]
```

By considering proton distributions, 3 subclasses can be distinguished. Do you want to see the protonated structures?Y

```
  CH
C |
H-C-OH
2 |
  CH
```

```
  CH
C |
H-C-OH
2 |
 CH2
```

```
 CH2
C |
H-C-OH
2 |
 CH2
```

[This fact alone is sufficient to consider a dehydration experiment, which is the experiment performed by the chemist when the work was originally done.]

[If desired, each of the three subclasses can be expanded

in turn. Only the expansion of the first class is shown (this class contains the correct structure).]

Do you want this feature to be further enlarged?Y (each subclass will be enlarged separately)

```
  CH
C |
H-C-OH
2 |
  CH
```

[PLAN can continue expansion of each subclass to search for further discriminatory features if requested. The results are omitted for brevity.]

present in 24 structures

```
  CH
C |
H-C-OH
2 |
 CH2
```

present in 49 structures

(end of report)

```
 CH2
C |
H-C-OH
2 |
 CH2
```

present in 15 structures

(end of report)
(continuing now with earlier report stage)
(end of report)
(continuing now with earlier report stage)
(end of report)
Do you want the program to check double bond systems?N

2.4 The Reaction Chemistry Program

During the past year we have made good progress in developing the reaction chemistry program, REACT, into a working tool for laboratory chemists. Two main areas of application are discussed in the subsequent sections. These areas and the examples included are currently in the process of appearing in the literature. Additional details can be obtained by referring to those papers when they appear. The first area of application (subsequent section) is the subject of a paper to appear soon in Tetrahedron. The second area is being written up for publication in the Journal of Chemical Information and Computer Science.

2.4.1 Studies in the Biosynthesis of Natural Products

Manual elucidation of structures arising from chemical reactions which may yield a large number of products via a number of complex, interrelated pathways is a difficult problem. Such reactions are, however, natural candidates for computer-assisted studies because the computer can easily record all intermediates and products as well as interrelationships among them. [22] [1,2] Examples of these reactions include carbonium ion rearrangements, reactions of free radicals and biochemical processes.

REACT is designed to carry out representations of chemical reactions on representations of chemical structures. Reactions, defined by the chemist using the program, are carried out in the synthetic direction as opposed to the retro-synthetic direction of programs for computer-aided synthesis. [3] In structure elucidation problems, the set of structures undergoing reaction is the current set of candidate structures for an unknown. It is clear, however, that the program can also be used effectively in following reactions of a single, known compound participating in a complex sequence of reactions. For example, we showed [22] that CONGEN together with REACT provides a convenient method for studying acid catalyzed rearrangements such as the conversion of tetrahydrodicyclopentadiene to adamantane. In that example, the complete set of isomers was generated by CONGEN. Subsequently, a one-step reaction carried out on each isomer afforded the complete rearrangement graph. An alternative method, similar to that discussed in subsequent sections, is to use a single isomer as a precursor. In the examples given in this work, a single precursor was subjected to repetitive application of a set of reactions.

---

[1] T. M. Gund, P. v. R. Schleyer, P. H. Gund and W. T. Wipke, J.Am.Chem.Soc. 97, 743 (1975).

[2] S. A. Godleski, P. v. R. Schleyer, E. Osawa, Y. Inamoto and Y. Fujikura, J.Org.Chem. 41, 2596 (1976).

[3] E.J. Corey and W.T. Wipke, **Science 166**, 178 (1969).

To demonstrate the utility of REACT we present two examples where a given precursor of known structure is subjected to an extended sequence of reactions. At each step in the sequence one or more reactions may apply to the products from the previous step. As will be shown in the sequel such an approach is especially well suited to problems involving the biosynthesis of natural products. A complete description of this work will appear shortly [22].

2.4.1.1 Generation of Biosynthetically Plausible Sterol Side Chains

Sterols are naturally occurring steroidal alcohols (usually 3-ols) which differ in the number and the position of methyl groups and the degree of unsaturation (present as a double bond or cyclopropyl ring). New sterols are frequently isolated in minute quantities from natural sources. Because of their structural similarities and the large number of different sterols present as a mixture in the same source (a recent paper [4] documents the isolation of ca. 50 sterols from one marine source) it is often difficult to separate them and to obtain pure compounds in quantities large enough for structure determination by conventional methods. Some structural assignments are based on biogenetic considerations, assuming that compounds from the same origin are related to each other through formation along the same biochemical pathway. This pathway can be a series of complicated chemical reactions which yield a large number of intermediates and products. It is difficult to follow manually such a series of reactions in order to explore all possible structural alternatives. To date, over 100 different 3-hydroxy sterols have been isolated, the majority of them based on the seven nuclear skeletons [5].

We use a method of combined gas chromatography/mass spectrometry (GC/MS) to analyze complex mixtures of sterols in a search for new compounds which may represent important biosynthetic intermediates. Part of this method involves research in interpretation and prediction of mass spectra. [23] We have used the REACT program as an additional tool to predict plausible structural candidates to guide both our manual and computer-based interpretations.

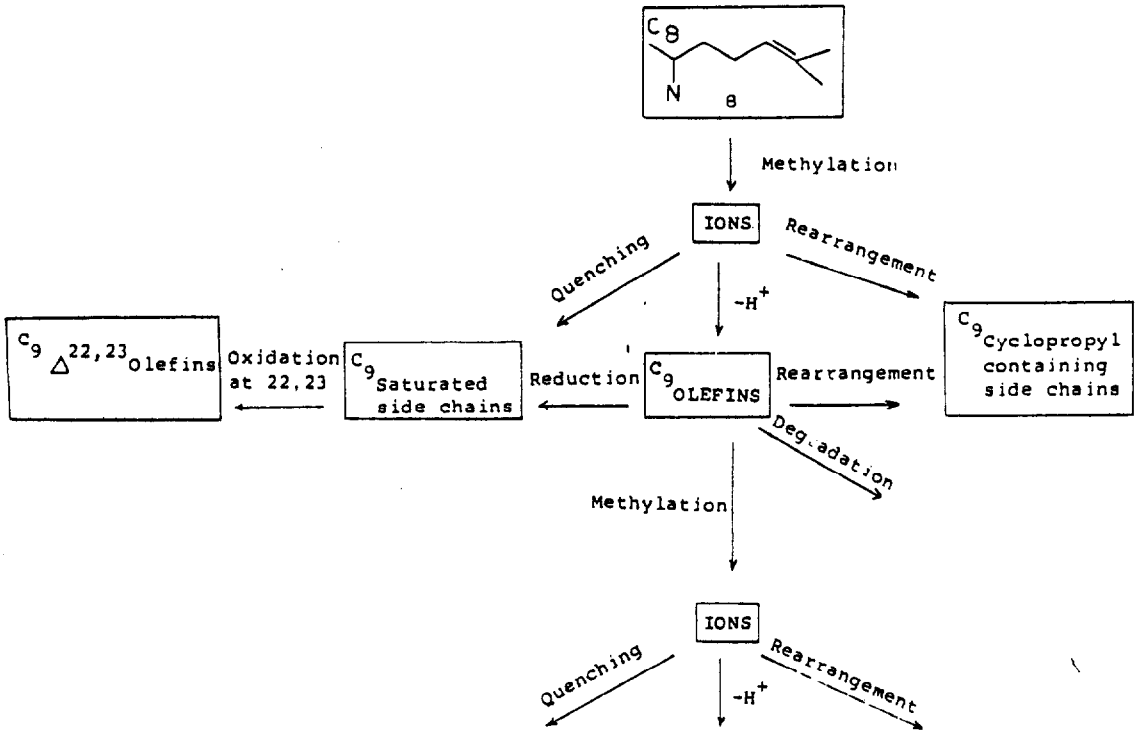
The set of reactions used in REACT to carry out possible transformations of sterol side chains have been suggested

---

[4] S. Popov, R. M. K. Carlson, A. Wegmann and C. Djerassi, Steroids 28, 699 (1976).

[5] C. Djerassi, R. M. K. Carlson, S. Popov and T. H. Varkony, in "Marine Natural Products Chemistry" by D. J. Faulkner and W. H. Fenical (ed.), Plenum : New York, N.Y., 1977, p 111.

previously [6] The precursor, (a 24,25 unsaturated side chain numbered 8 at the top of the following chart) the order of application of the various reactions and the classes of products which result are shown in the following chart. The sequence of reactions consists of repetitive application of the following steps:



[6] E. Lederer, Quart.Rev.,Chem.Soc. 23, 453 (1969).

1) Methylation. C-methylation of a double bond. In nature this reaction occurs via the ylide of S-adenosylmethionine. This reaction is constrained for general application later in the sequence to forbid the sterically unfavorable methylation of tetra-substituted double bonds.

2) The carbonium ion obtained by the alkylation can undergo several reactions:

a) proton elimination and formation of a double bond; b) cyclization to form a cyclopropyl system with subsequent elimination of a proton; c) quenching to form saturated side chains.

3) The olefin is allowed to undergo several additional reactions:

a) reduction to form a saturated side chain; b) rearrangement to a cyclopropyl system; c) degradation to shorter side chains via loss of allylic methyl groups; d) methylation to produce longer side chains.

Constraints on reactions of the olefin included

a) subsequent migration of the double bond is not allowed; b) olefins obtained by degradation are allowed to undergo only one step of methylation.

4) Subsequent oxidation of saturated side chains proceeds to form a new double bond at C-22,23, a mechanism proposed by Knapp, et al. [7] This set of reactions was applied sequentially a total of three times. Thus, side chains possessing from seven to eleven carbon atoms are accessible by this sequence.

Results. A numerical summary of results is presented in our Table below. The table is organized by summarizing the side chains produced by the different biochemical pathways. The only known, naturally occurring C7 saturated side chain was correctly predicted by REACT. Three C7 unsaturated side chains were predicted. Two of these three exist in nature. In the C8 series five unsaturated side chains out of 12 predicted are observed in nature. For the longer side chains, more are possible but fewer are observed. For example, only one out of the 76 predicted C11 side chains has so far been found in nature.

---

[7] F. F. Knapp, J. B. Greig, L. J. Goad and T. W. Goodwin, J.Chem.Soc.,Chem. Comm. 707 (1971).

Number of C in side chains	SATURATED				OLEFINS					CYCLOPROPANES		
	A	B	E	Nature	A	B	E	F	Nature	C	D	Nature
7	-	1	-	1	-	2	-	1	2	-	-	-
8	1	2	-	1	1	6	3	2	5	-	-	-
9	1	7	-	1	4	13	6	4	4	2	4	-
10	3	12	-	4	13	17	19	8	6	8	11	1
11	8	-	8	1	31	-	37	8	1	17	21	1

A methylation only.
B methylation followed by degradation only.
C rearrangement of carbonium ion.
D rearrangement of olefin.
E degradation followed by methylation only.
F oxidation of saturated side chains at 22,23 position.

Table I. Number of Side Chains Produced by Different Pathways.

The total number of sterols which obey our biosynthetic constraints is 1778. This number is manageable by techniques of computer-assisted structure elucidation. Separating the structures by molecular weight reduces considerably the number of candidate structures which must be considered in a given problem. Thus, in a GC/MS experiment the maximum number of structures we have to consider is not larger than 264 (the number of isomers with empirical formula $C_{29}H_{48}O$. Any additional spectroscopic or chemical data reduce this number still further. For other molecular weights the number of possibilities is considerably fewer. Structural information from the mass spectral fragmentation pattern of the molecule may leave only a small number of possibilities from which to choose.

2.4.1.2 Elucidation of Biosynthetic Pathways

Elucidation of biosynthetic pathways can be accomplished in several ways, including for example co-occurrence of structurally related compounds or use of mutant organisms which accumulate

intermediates. [8] These methods usually leave the structures of intermediates and/or the details of the biochemical pathways open to question. More detailed experiments are required to establish rigorously reaction pathways from precursor to product.

Isotopic labelling experiments are capable of providing additional detail through synthesis of labelled precursors followed by incorporation of labelled substrate and determination of the labelling pattern of the products of biochemical transformation. The incorporation of labelled precursors into desired products is generally low and elucidation of the labelling pattern in minute amounts of product is difficult. Thus, these experiments are generally time consuming and costly. They can be complicated by the existence of different biochemical pathways, some of which yield products with the same distribution of isotopic labels. Therefore, care must be used in designing such experiments. It is important to select a labelled precursor that will allow one to distinguish among most of the possible pathways, and that will lead to a product with labels distributed in easily detectable positions. Manual methods are often insufficient to determine all the theoretically possible pathways when the number of possible pathways and the number of intermediate structures is very large. However, this type of problem is easily managed by REACT, which can accurately and systematically monitor transformations of the precursor into products, follow the isotopic labels throughout a reaction sequence and detect the formation of equivalent structures and labelling patterns. We stress that this is not an exercise in "paper chemistry", but a systematic way to investigate all the possible aspects of a proposed experiment before devoting valuable time and resources to an experiment which leads to ambiguous results.

An example which illustrates our method is the exploration of biosynthetic pathways leading to formation of a family of fungal metabolites [9] The complete paper [22] describes our results in detail. Briefly, use of REACT enabled us to: 1) verify proposed pathways and suggest alternatives; 2) demonstrate how different patterns of isotopic labelling lead to unambiguous assignment of pathways for certain molecules; and 3) demonstrate that several pathways are possible for certain other fungal metabolites, pathways which would not be differentiated by proposed labelling schemes.

2.4.2 Applications to Structure Elucidation

The first version of REACT and its applications were

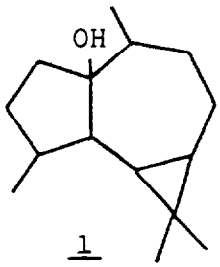
---

[8] J. D. Bu'Lock, "The Biosynthesis of Natural Products", McGraw Hill, New York, N.Y., 1965, p.94.

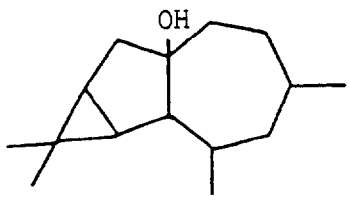
[9] G. A. Cordell, Chem. Rev. 76, 425 (1976).

described previously [22]. Subsequently, the structure of the program was revised significantly to include commands and internal operations which more closely parallel laboratory procedures. The new version has been described briefly and some applications of REACT to mechanistic problems have been discussed [24]. In subsequent sections we describe the REACT program in detail, together with an example of the application of the program to a structural problem.

To demonstrate the application of REACT we choose an example which illustrates some (but not all) aspects of the use of REACT in a structure elucidation problem. A contrived example might illustrate many of the other features and subtleties of the program, but would not be as meaningful chemically. The example involves a dehydration reaction (see reaction definition) applied during the course of elucidation of the structure of palustrol (1)[8]. Structural features of the products were powerful constraints on the identity of the compound. This problem was solved prior to the existence of the REACT program.



1



2

We pick up the example at the point at which the reaction was applied in the laboratory. This example is of interest because it represents a case where direct translation of observations on products back to structural constraints on the starting materials is difficult. Using REACT, expression of structural information is straightforward and logical. The laboratory reaction, separation and key structural information are summarized below. The starting materials, in a flask called STRUCS, are the candidate structures for palustrol (1).

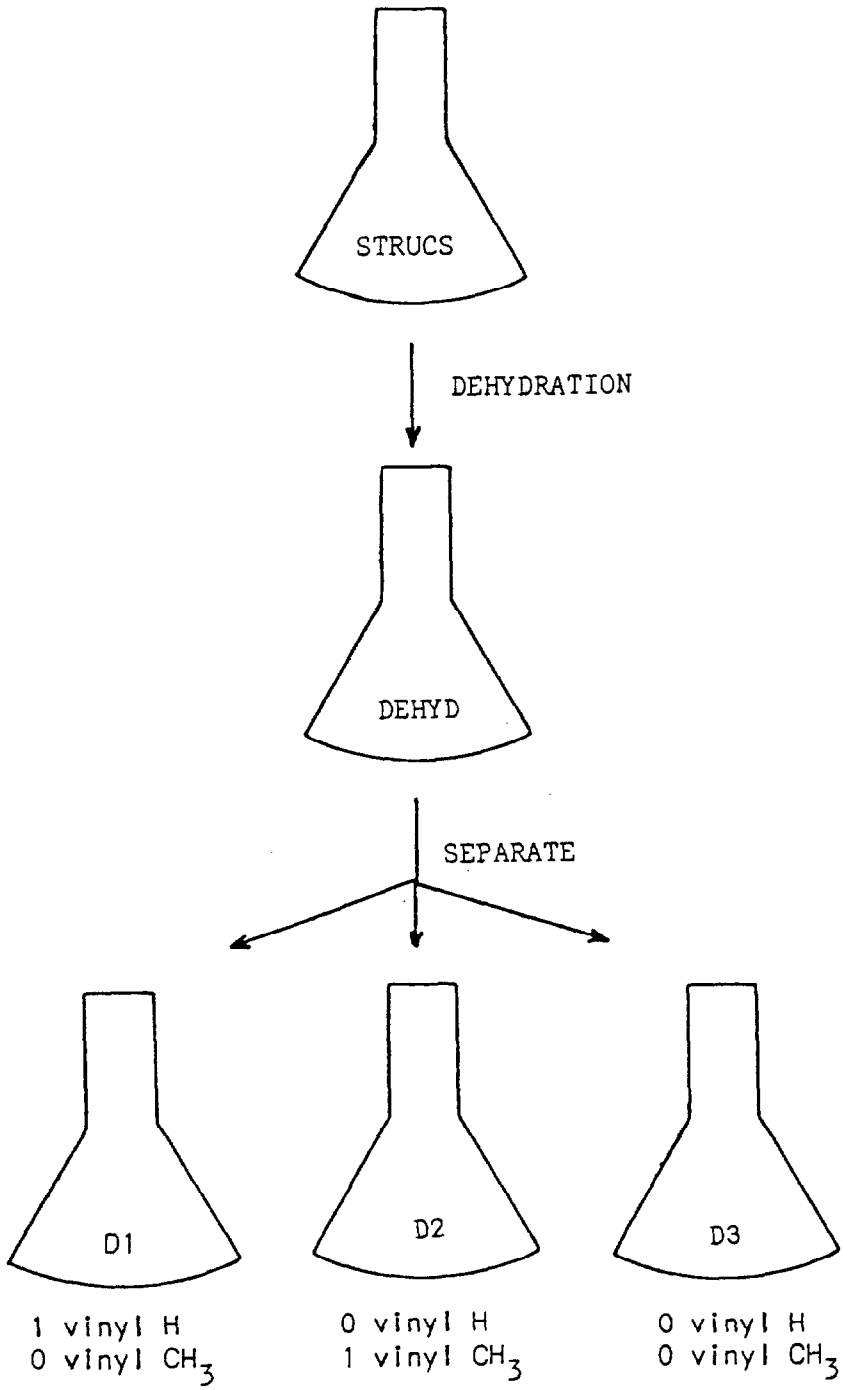


Figure 1. Diagram of REACT's Separation of CONGEN Structures with Respect to Dehydration.

Consideration of all available spectroscopic data had reduced the problem to a set of 88 candidates prior to carrying out the dehydration reaction. The contents of the flask STRUCS were dehydrated and the products placed in a flask called DEHYD. Separation of the reaction mixture yielded three products, placed in flasks D1, D2 and D3. The numbers of vinyl protons and vinyl methyl groups detected by H NMR for each product are summarized in Figure 1.

2.4.2.1 The Reaction Tree

The reaction tree is a representation of the sequence of laboratory procedures (reactions and separations) to which precursors and their products have been subjected. Formally, it consists of named flasks and their interrelationships in the form of reaction names and separation steps. If there are multiple precursors (i.e. more then one structure in a flask), as in the example, each is allowed to react, independently, resulting in a data structure internal to REACT which records the reactions of each structure separately. The chemical meaning of multiple structures in the starting material flask STRUCS is that the exact identity of the compound is not known; its structure is represented by one of all the possible structures in the flask. If the flask was created via a reaction(s), the structures represent the collection of all products from all precursors where, again, the identity of each of the products in the laboratory application of the reaction is not necessarily known. In our representation, an example of which is shown in Figure 1, flasks which could possess multiple structures, such as multiple candidates for an unknown, are depicted as containing all structures, and all possible products appear lumped together in a product flask. The dehydration reaction applied above (see Table III) is summarized in Fig. 2.

```
STRUCS=88
|
*DEHYDRATION->DEHYD=241
```

Figure 2. Result of Dehydration in REACT

This figure is interpreted to mean that the 88 candidate structures, any one of which could be the true unknown in the flask STRUCS, yield a total of 241 possible products, all associated with the flask DEHYD. Confusion related to this presentation can be avoided by remembering that the internal representation is effectively n copies of the reaction tree where n is the number of precursors in the flask STRUCS, or 88 for the

example of Fig. 1 For example, one such copy encodes the information about the conversion of 3 to 4a - 4c.

In our example we discuss only a single reaction. In general, however, the reaction tree can be of arbitrary complexity. Several different reactions can be applied to aliquots of a precursor (whether it be an original starting material or a product of a previous reaction). In addition, an extended sequence of reactions can be carried out. Thus, the reaction tree can grow arbitrarily in width and depth.

2.4.2.2 Separation

A flask obtained by reaction can contain a mixture of products. A single precursor can yield multiple products in three ways in a reaction: 1) presence of multiple reaction sites, each yielding a different product; 2) multiple reactions; and 3) cleavage reactions where all fragments are isolable. The usual laboratory step subsequent to reaction is separation of the products. Thus, REACT has a SEPARATE command which allows the chemist to express to the program his laboratory observations on performing the separation. The number of products obtained on separation is a constraint on the identity of the starting material, and is information useful in applications of REACT to structural problems. The separation requires placement of each separated product into a designated or named, flask (Table II).

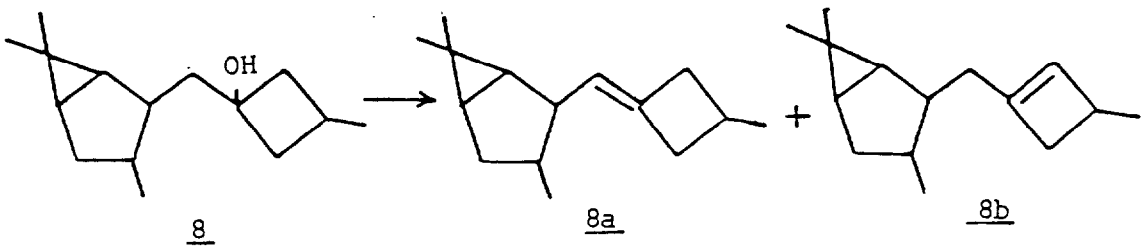
Table II. The Dialog with REACT on Separation of Contents of Product Flask
DEHYD into Flasks D1, D2 and D3

Command	Comment
SEPARATE	Enter separation mode
NAME OF FLASK TO BE SEPARATED:DEHYD	Select product flask
NEW FLASK NAME:D1 NEW FLASK NAME:D2 NEW FLASK NAME:D3	Select names for flasks for three separated products
NEW FLASK NAME:	No other flasks
MAXIMUM NUMBER OF ADDITIONAL PRODUCTS:0	No additional products in the tar flask
210 STRUCTURES SURVIVED SEPARATION	Results
BEGINNING RAMIFICATION...DONE	Implications of separation
	Return to REACT

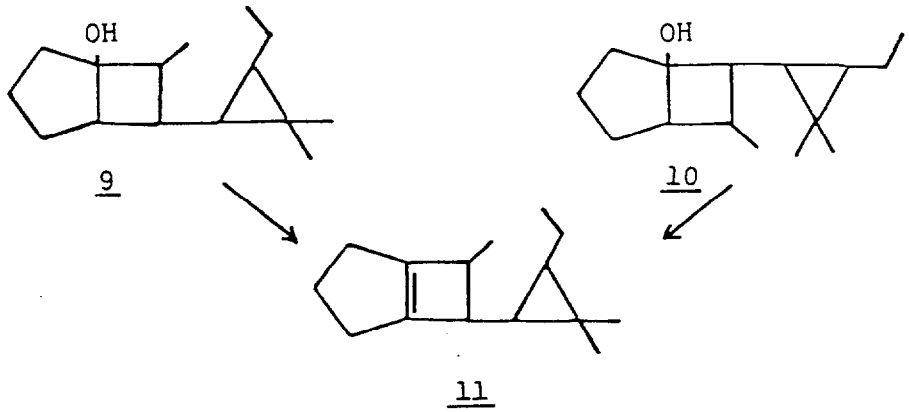
It is characteristic of many laboratory reactions that an unspecified, perhaps large, number of additional products are obtained, some legitimate, but at low concentration, others from side reactions which may not be incorporated in the definition of the reaction used in REACT. The chemist using REACT must base his use of the SEPARATE command on his own evaluation of the reaction applied in the laboratory. Selection of a named flask in which to place a separated product implies that the product so separated arose from the named reaction, and not from some other unspecified reaction. However, to accommodate the fact that the reaction may have been incomplete or side reactions may have occurred, additional products can be specified to be in a "tar" flask associated with each set of separated products. On separation, the new flasks each contain one unique product, whose identity is not known. The structure of the product must be one of the structural possibilities associated with the flask. However, the structures in the "tar" flask, (or in any flask prior to separation) can be a mixture of products, where each product in the mixture may be represented by several structural possibilities.

The dialog to establish separated products and a tar flask with REACT is summarized in Table II. In the laboratory, separation yielded three products (Fig. 1). In this example we choose to specify exactly three products by selection of three flasks to receive the products, D1, D2 and D3, and no other.

The fact that three products, all assumed to arise from the dehydration, were observed is a constraint on the identity of the starting material in the flask STRUCS. Those structural possibilities (according to CONGEN) for palustrol which would yield only two products (e.g., 8, to yield 8a and 8b) can be rejected independently of the identity of the products, while those structures which yield three products on dehydration remain under consideration (e.g., 1 and 2) until additional data on the identities of the products are gathered and specified to REACT (see subsequent section).



8 8a 8b



The reaction tree which results from the separation (Table II) is shown in Figure 3.

```
STRUCS=72
|
*DEHYDRATION->DEHYD=210-s-|D3=210
                          |
                          |D2=210
                          |
                          |D1=210
```

Figure 3. Results of Separation in REACT.

The reduction in the numbers of structures in flasks STRUCS and DEHYD (compare Fig. 3 to Fig. 2) results from the implications, or ramifications, of the statement on separation. REACT has a record of how many products are obtained from each structure and the identities of each precursor and product. It can eliminate automatically from further consideration precursors which yield an undesired number of products. If three products are observed, as in the example, only 72 of the original 88 structures remain as candidates. Sixteen of the structures yielded, by the computer program, other than exactly three products and were therefore removed from consideration as candidates. The products of these sixteen structures are also removed from the product flasks, resulting in a decrease in the number of structures in DEHYD from 241 to 210. The remaining 210 structures are not exactly three times 72 because several candidates yield equivalent products. For example, the dehydration of both 9 and 10 yields, among other products, 11.

As mentioned previously, duplicate structures are detected and removed for efficiency, except in mechanistic reactions.

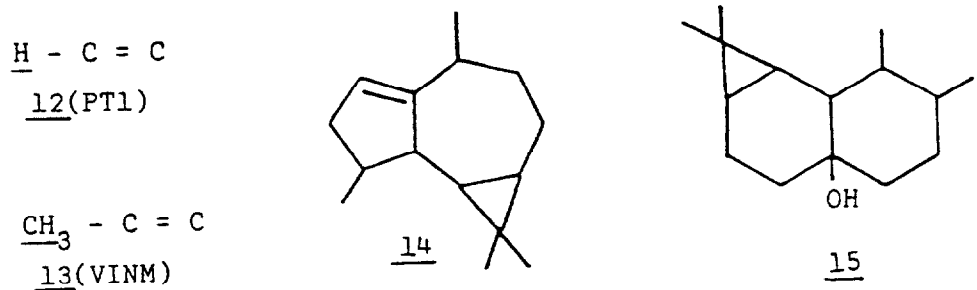
What of the contents of the flasks D1, D2 and D3? Up to this point, no statements about the structural identity of any product have been made, paralleling the laboratory events of, first, separation, and, later, gathering of data on the products. Thus, any of the 210 products in DEHYD might be in any of the flasks D1 - D3 (see Fig. 3, where all 210 products remain allocated to D1 - D3). Stated at the level of internal representation in REACT (see also above discussion), where the original structures are represented individually, each structure (in STRUCS) yielded three products, any of which might be in any flask. Subsequent operations will perform the appropriate allocations of structures to flasks.

Details of the internal representation and the algorithm which performs ramification after SEPARATE and PRUNE (see below) are given in a separate publication. This algorithm is responsible for determining legal allocations for structures to flasks throughout the reaction tree whenever the tree is modified in any way.

2.4.2.3 PRUNE - Expression of Constraints on Products

In laboratory procedures, the next step would be to collect data on the product in each flask. Structural information gained represents constraints not only on the identity of the products, but also on the identity of the precursor and its precursor and so forth throughout an entire reaction sequence. REACT allows structural statements to be made as constraints on the contents of any flask in a reaction tree. The command to express constraints is PRUNE (a word which is jargon but does carry with it the concept of trimming the reaction tree and also corresponds to the same command in CONGEN.

Substructural constraints can be obtained from a file or defined by the chemist as required, using EDITSTRUC. In our example, the product in one of the flasks (D1) was observed according to H NMR analysis to possess one vinyl proton and no vinyl methyl groups. These substructures, PT1 (12) and VINM (13), respectively, were defined and the substructures supplied to PRUNE.



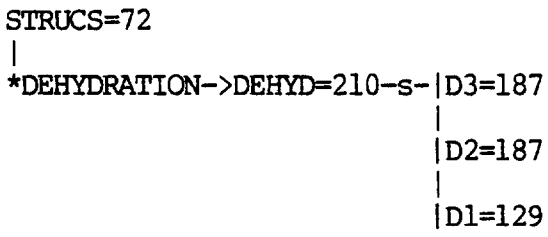


Figure 4. Application of PRUNE in REACT.

The reaction tree which results on application of PRUNE is shown in Figure 4. There remain 129 structures which could be in the flask D1. The number of structural candidates (72) has not been reduced, implying that all 72 can yield at least one structure possessing one vinyl proton and no vinyl methyls. Some candidate structures can yield more than one product which obeys these constraints and might therefore be in D1, resulting in 129 rather than only 72 structures in that flask. For example, 2 yields two products obeying the constraints; either could be the product observed in D1. However, for structure 1, only one of the products (14) is a legal structure under the constraints; that structure must be in flask D1.

If one product is forced to be in a certain flask it can be in no other flask. Thus, the number of dehydration products which could be in D2 and D3 decreases from 210 to 187 (compare Figs. 3,4). Obviously, with a more complex reaction tree, such logical decisions become complicated. REACT determines allowable allocations automatically.

Flask D2 contains a product which possesses no vinyl protons and one vinyl methyl group (Fig. 1). Constraining the contents of D2 with this structural information results in the allocation summarized in Figure 5.

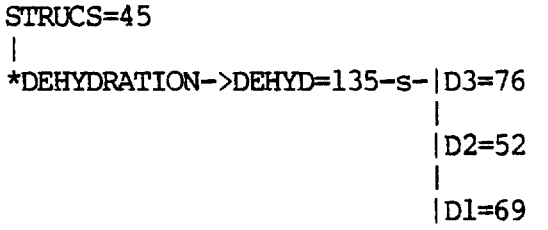


Figure 5. Results of Constraining Contents of Flasks in REACT.

Now the number of candidate structures in STRUCS is reduced to 45, implying that there are 72-45=27 structures which cannot

yield a product distribution which satisfies the structural constraints placed on both flasks D1 and D2. An example is 12, which, although it yields at least one (two) products satisfying the constraints on flask D1, yields no products satisfying the constraints on flask D2. It is therefore discarded as a candidate structure. At the same time, any products of discarded structures (and precursors in a more complex tree) are removed from DEHYD and flasks D1 - D3.

Application of the constraints on flask D3 (Fig. 1), that the product contained therein possess neither a vinyl methyl nor a vinyl proton results in the reaction tree shown in Figure 6. Now only fourteen structural candidates remain, and from the allocation of products to flasks (Fig. 6a) each yields three unique products. Each of the structural candidates was tested for the presence of exactly two secondary methyl groups; the reaction tree of Figure 7 results.

Previously, translation of the results of the dehydration into a substructure used to test the 88 candidates reduced the number of candidates to 22, rather than 14 (Fig. 6a).

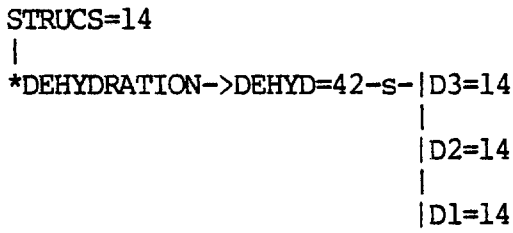


Figure 6. Further Application of Constraints.

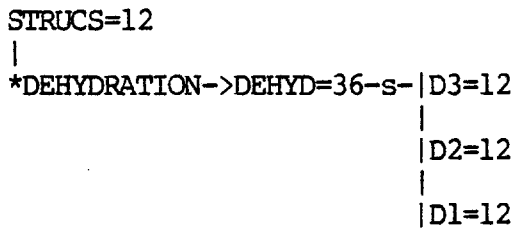


Figure 7. Constraining Contents of Flasks Still Further.

The substructure used was correct, but incomplete in that eight structures which obeyed the substructural constraint could not yield the observed products. Through use of REACT, structural information can be applied directly to the structures of potential products without the necessity of translating observations back to the precursors.

2.4.3 Utilities

We discuss the utilities briefly here not because they are critical to understanding the method but because they are an essential part of the interactive nature of REACT.

1) **Displaying Reaction Tree.** Examples of reaction trees in Figures 2-6 illustrate the format in which the reaction sequence can be observed. The DISPLAY1 command allows the chemist to view selected portions of the tree, i.e., one named flask together with any separations or reactions performed on that flask.

2) **Drawing Structures.** The structures (or any subset) in any selected flask can be drawn. To check numbering of atoms, particularly in the use of MREACT, structures can also be drawn with structure numbers (NDRAW).

3) **Determining Structural Relationships.** Relationships between precursors and products can be obtained using the PARENTS and PRODUCTS commands. A report can be obtained for all or selected structures in a flask, either to summarize precursors which led to a structure (PARENTS reports flask and structure number of every parent of every structure) or products of all or selected structures (PRODUCTS reports flask and structure number of every product of every structure). These commands were used to examine the reaction tree in the example to determine relationships among structures presented in the text.

4) **File Manipulation and Other Commands.** These utility commands allow a chemist to save and restore problems or portions thereof at will, thereby maintaining a computer-based "lab notebook" of his operations. Other commands simplify the reporting of problems and subsequent improvement of REACT and correction of errors. CHECKPOINT and UNDO are useful when the chemist wants to explore the consequences of a separation or pruning and still return to his previous reaction tree if desired.

2.5 Mass Spectral Prediction and Ranking

2.5.1 Predicting Spectra Using MSRANK and the Half-Order Theory

The MSRANK program has been incorporated as part of CONGEN, but is not yet available for general use by outside persons accessing CONGEN. We have during the past year been giving the program some extensive tests to determine its scope and limitations. We have studied the following classes of compounds (all closely related to current research problems): 1) marine sterols; 2) substituted pregnanes; 3) aliphatic and aromatic esters; and 4) macrolide antibiotics.

We conclude that MSRANK is a powerful filter for eliminating from further consideration structures which cannot yield the observed mass spectrum for an unknown by "reasonable" fragmentation pathways. The greater the structural diversity of isomeric candidates for an unknown, the better the performance of MSRANK in focussing in on the correct structure. When the structures are quite similar, for example when they have been constructed from the same set of superatoms and few remaining atoms, the ranking by MSRANK is quite similar (as one might expect). When this situation occurs, the chemist must still consider the top 10 - 50 percent of the structures as possibilities, depending on the distribution of scores.

We have added an explanation feature to MSRANK. Upon request the program prints a list of peaks in the observed spectrum which have different "reasonable" explanations. for different candidate structures. Based on this information the chemist can accept the ranking or change the parameters which define his theory of fragmentation to obtain a different ranking. This procedure helps detect and reduce the plausibility of "nonsense" fragmentation processes.

2.5.2 Prediction Using Fragmentation Rules Supplied by Chemists

When the candidate structure is known to belong to a previously investigated class of compounds, then we can use additional information to predict a more precise mass spectrum. This information is in the form of specific fragmentation rules. These rules are described by a subgraph, a break (or cleavage) and related hydrogen or neutral transfers, intensity ranges associated with rules and a parameter describing the confidence in a rule. We are working on a program which allows the user to enter rules defining his theory of mass spectral fragmentation. A computer session for entering rules which describe fragmentation of ring D in 17-substituted steroids is presented below to convey the nature of a fragmentation rule and associated parameters.

```
@<wcw>dendrl                        <begin program>
using <LISP>CARHART.SAV;70702
<WCW>DENDRL.SAV;8 created 26-JAN-78 06:06:39
what do you want to do? : CRF
create user rule file.

new rule set.
 = ?                                <query for options>
one of the following:
RESTORE  ENTER  DELETE  SHOW  SAVE  QUIT  ??
 = ENTER R1                         <enter rule named "R1">
enter rule:
 := SHOW                            <query rule>
```

```
working space:
rule R1                                  <not defined yet>
 := ?                                    <query for options>
one of the following:
CLEAR  FETCH  NAME  GRAPH BREAK PEAKGROUP DRAW SHOW ADD QUIT
??
 := GRAPH
entering editstruc.                      <use EDITSTRUC for defining
                                          subgraphs>
(NEW STRUCTURE)
>RING 5
>BRANCH 5 1 1 1 1 1 2 1
>NDRAW

R1
    3   9
   / \ /
  4   2
  |   |                                  <define and draw subgraph>
  5---1
 /   / \
6   8   7
>DONE
(R1 DEFINED)
                                         <specify cleavage by naming
                                          bonds cleaved between
                                          numbered atoms>
 := BREAK (1 2)(5 4)
 := PEAKGROUP                            <specify what peaks are
new list of peak groups.                  produced by the cleavage>
 ::= ?                                   <query for options>
one of the following:
DELETE  CLEAR  FETCH  NAME  TRANSFERS  SIGNIFICANCE  INTENSITY
SHOW  ADD  QUIT  ??
 ::= NAME P1                             <name peakgroup>
 ::= T H -1                              <include loss of hydrogen>
 ::= SHOW
working space:
peak group P1
(TRANSFERS -1)
 ::= INTENSITY 80                        <assign relative intensity
 ::= SIGNIFICANCE 90                      and plausibility>
 ::= ADD
P1 included in PEAKGROUPS
next:
 ::= NAME P2                             <name P2>
 ::= T H -1 H2O -1                       <accompanied by loss of
 ::= SH                                   water and hydrogen>
working space:
peak group P2
(TRANSFERS :-H-H2O)
 ::= I 50                                <assign relative intensity
 ::= SI 60                                and plausibility>
 ::= SH
```

```
working space:
peak group P2
(TRANSFERS :-H-H2O)
(INTENSITY 50)
(SIGNIFICANCE 60)
 ::= AD
P2 included in PEAKGROUPS
next:
 ::= Q
 := SHOW
working space:
rule R1                                  <summarize rule R1>
show subgraph drawing? Y/N. : N
show connection table? Y/N. : N
(BREAK (1 . 2) (5 . 4))
peak group P1
(TRANSFERS -1)
(INTENSITY 80)
(SIGNIFICANCE 90)
peak group P2
(TRANSFERS :-H-H2O)
(INTENSITY 50)
(SIGNIFICANCE 60)
 := ADD
R1 included in RULES                     <add R1 to list of rules>
next:
 := N R2                                 <define R2>
 := G                                    <etc....>
```

Applying these rules to a set of candidate structures produces a predicted spectrum. This predicted spectrum is different from the one created by MSPRED or MSRANK, and more closely resembles an observed spectrum. The peaks in this predicted spectrum have different intensity values, and the density of the spectrum i.e. the number of peaks per mass range is smaller. This minimizes the number of incorrect predictions and makes the entire predicted spectrum more closely related to an observed spectrum of an unknown compound. We are also working on a program to plot predicted spectra. This will be useful for visual comparison of plotted observed spectrum against a predicted one.

The next step is to explore ways to compare a predicted and an observed spectrum. We are experimenting with different ranking functions (see section 4) and developing a program which will allow the user to define in a simple mathematical equation his individual ranking function. The problem of ranking candidate structures based on spectrum comparison is closely related to the problem of library search. In our case, however, we do not have authentic spectra of our structural candidates in most instances. The density of a predicted spectrum for a candidate is quite low

because we do not attempt to predict the complete spectrum. Rather, we predict major fragmentations. This fact must be taken into account in designing a function to rank candidates based on comparison of their predicted spectra to that of the unknown.

2.6 Molecular Ion Determination

The original MOLION program [10] was based upon the postulate: "There exists at least one SECONDARY LOSS in a spectrum that will match a PRIMARY LOSS from the molecular ion irrespective of whether the molecular ion is present in the spectrum."

Given this postulate, then one method of generating candidate masses for a molecular ion (M+) is to identify all possible secondary losses apparent in a spectrum, and then to add each of these losses to the masses of those ions observed in the high mass region of the spectrum. This, together with some refinements, was the basis of the original MOLION program. The most important of these refinements were:

(i) "PLANNING" i.e. the filtering of the set of apparent secondary losses against a table of "bad losses" (containing chemically implausible values like 9 amu and 23 amu), thus reducing the number of initial candidate M+s.

(ii) "TESTING". An acceptable candidate M+ had to be greater than, or equal in mass to the highest mass ion observed, and none of its immediate losses to observed ions could be in the list of "bad losses".

There are, however, a number of problems with the algorithm used in MOLION. The most crucial problem is that the algorithm requires good spectra! Impurities such as column bleed or co-eluting minor components can result in ions that would constitute bad losses --- causing the rejection of distinct and well supported molecular ions recorded in the spectrum. Further, the program did not allow the user to modify the "bad loss" set, nor to have access to the molecular ion scoring mechanisms. These scoring mechanisms incorporated a considerable measure of class dependency. Thus when testing a candidate M+, the program could modify the score associated with the M+ by the intensity combination formula: e.g. a mass difference of 101amu between the candidate M+ and an observed ion resulted in a 1.8 times increase in that M+'s score whereas a mass difference of 2 or 16 reduced the score by 85 per cent and a difference of 44, 56, 60 or 72 reduced the score by 25 per cent.

---

[10] R.G.Dromey, B.G.Buchanan, D.H.Smith, J.Lederberg and C.Djerassi. "Applications of Artificial Intelligence to Chemical Inference. XIV. A General Method for Predicting Molecular Ions in Mass Spectra." Journal of Organic Chemistry40770 (1975).

In devising the new version of the molecular ion program, an attempt has been made to recognize and overcome some of these problems. The resulting program has the following new characteristics:

1) The user has complete control of all aspects of the candidate evaluation procedures; these evaluation procedures being defined in terms of conventional chemical concepts.

2) The scoring algorithm allows for the separate accumulation of evidence supporting and disconfirming a particular candidate mass for M+. Simple yes/no tests, like MOLION's "bad losses", are not used. In this way, the program is made a little more tolerant of impurities in the spectrum, etc.

The basic algorithm remains: candidate M+s are generated and then ranked according to whether they are of the expected parity, show chemically favorable or unfavorable losses etc.

The candidate generation procedure allows for candidates in the mass range I-115 to I+115 where I is the highest mass observed ion. Any ion in this region is a candidate, as is any mass that can be obtained by adding an apparent neutral loss to the mass of an observed ion. The apparent neutral losses are simply the mass differences between all pairs of ions in the spectrum. No chemical information is used at this stage. The set of apparent neutral losses is not filtered against any "bad loss" set; consequently, the set will contain many losses that cannot correspond to any conceivable chemical fragmentation (e.g. 9amu). This initial candidate generation procedure does contribute to the scores of candidate M+s; a candidate is scored proportional to (i) the number of ways it can be generated, and (ii) the importance accorded to the losses involved.

Once all M+ candidates have been generated, each is scored according to rules describing molecular ion properties. These rules include tests on parity, testing that the candidate M+ ion is the most intense ion in the region M-2 to M+4, determining whether there are ions observed at higher masses etc.

Finally, rules defining chemically important fragmentation processes are applied. These rules can be general in form, e.g. just specifying that losses such as 7, 9, 22 amu etc are chemically implausible and so, if ions are observed that would involve such losses from a candidate M+ then that candidate has to be down rated. In addition, it is possible for the chemist to specify class specific rules. Thus, if the chemist knows that loss of methyl groups and water is characteristic of the compounds he is analyzing, then he may specify that 15 and 18 are good losses which, if observed, should increase the score of a candidate M+.

The scoring scheme uses the Confidence Factor model of the MYCIN program [11] This Confidence Factor (CF) model is intended for situations where a proper Bayesian statistical approach is inappropriate (because the requisite a priori and conditional probabilities are not known). The CF model simply requires that the chemist be able to express, in semiquantitative form, ideas like:

> "the occurrence of an ion with the mass of a particular candidate M+ strongly increases my belief in that candidate being correct."
>
> "the observation of an ion that could be due to loss of H2O from a candidate M+ slightly increases my belief in that candidate."
>
> "the occurrence of ions at masses higher than a candidate M+ greatly increases my disbelief in that candidate."

Separate measures are kept of the total evidence supporting and opposing each hypothesis. These are the measures of belief in a hypothesis given some evidence (BF(h,e)), and the measure of disbelief in the hypothesis (DBF(h,e)). The overall confidence in a hypothesis is given by the difference of these measures:

$$CF(h,e) = BF(h,e) - DBF(h,e)$$

The range of values allowed for the measures BF and DBF is from 0 (corresponding to no evidence) to 1 (proof); thus, the CF value for a hypothesis can range from -1 (total disbelief) to 1 (total acceptance).

As additional evidence is found that supports some hypothesis, the measure of belief in that hypothesis increases asymptotically to 1:

$$BF(h,e1\&e2) = BF(h,e1) + (1 - BF(h,e1))*BF(h,e2)$$

(it is implicit that e1 and e2 are independent). A similar formula defines how the accumulation of negative evidence causes the disbelief to increase asymptotically.

In many cases, there may be uncertainty about the premise of some rule for scoring candidate M+s. The MYCIN model allows rules to be used with reduced strength when there is uncertainty about the premise. Thus, given a rule of the form

> "if a candidate M+ can be generated by adding a known neutral loss to some observed ion's mass then my confidence in that M+ is increased by 0.1"

---

[11] E.H.Shortliffe and B.G.Buchanan. "A Model of Inexact Reasoning in Medicine." Mathematical Biosciences,23,351 (1975).

then if one were only 0.6 confident that some apparent neutral loss was significant in a spectrum this rule would be used to support a candidate M+ with a strength of belief of 0.06 (product of confidence in premise of rule and strength of inference if one were certain that premise was true).

The chemist's control over MOLION's M+ evaluation scheme is expressed through the following parameters:

1) Parameters that influence the preference for even/odd mass candidate M+s.

2) Parameters expressing the importance accorded to a candidate M+ actually being observed in the recorded spectrum.

3) Significance of ions above a candidate M+.

4) Chemically significant neutral losses.

5) Identifying secondary losses from the recorded spectrum.

6) Use of H+ transfer to add extra losses to the set of apparent secondary losses.

The following example shows the set of rules used to process some example sterol spectra:

IF the ratio of even mass ions / odd mass ions exceeds 0.50 THEN: there is evidence for Nitrogen presence and belief in all odd mass M+s is increased by 0.10 ELSE: by default disbelief in odd mass M+ is increased by 0.20

IF the ratio of accumulated intensities of even and odd mass ions exceeds 0.50 THEN: there is evidence for Nitrogen presence and belief in all odd mass M+s is increased by 0.10 ELSE: by default, disbelief in odd mass M+ is increased by 0.20

IF a candidate even mass molecular ion is in the recorded spectrum, THEN: belief in that candidate is increased by 0.50 ELSE: disbelief in that candidate is increased by 0.30

IF a candidate odd mass molecular ion is in the recorded spectrum THEN: belief in that candidate is increased by 0.25 ELSE: disbelief in that candidate is increased by 0.25

IF ions occur above M+4 for some candidate molwt M THEN: disbelief in that candidate is increased by 0.20

IF the candidate molecular ion M is not the most intense in the range M-2 to M+4 THEN: disbelief in that candidate is increased by 0.40

The following neutral losses are held to be chemically significant and confirm belief in a candidate M+

15	0.50
18	0.50
31	0.20
33	0.25

The following neutral losses should not occur. If such a loss is implied by a candidate M+ then disbelief in that candidate is increased by the specified amount

3	0.05	26	0.40
4	0.10	28	0.10
5	0.40	30	0.10
6	0.40	34	0.40
7	0.40	35	0.40
8	0.40	36	0.40
9	0.40	37	0.40
10	0.40	38	0.40
11	0.40	39	0.40
12	0.40	40	0.40
13	0.40	44	0.10
14	0.40	45	0.10
16	0.10	46	0.10
19	0.10	47	0.10
20	0.40	48	0.10
21	0.40	49	0.10
22	0.40	50	0.10
23	0.40	51	0.10
24	0.40	52	0.10
25	0.40	53	0.10
		60	0.10
		61	0.10
		62	0.10

Each time that it is observed in the spectrum, belief in the reality of an apparent 2ndry loss is increased by 0.05

Belief in a candidate M+ is increase by 0.10 each time that it is generated by adding the mass of a known 2ndry loss to an observed fragment ion.

If loss of I amu has been identified as a 2ndry loss, but loss of I+1 amu was not apparent Then loss of (I+1) amu can, with confidence 0.50, be added to the set of losses used for M+ generation.

If loss of I amu has been identified as a 2ndry loss, but loss of I-1 amu was not apparent Then loss of (I-1) amu can, with confidence 0.50, be added to the set of losses used for M+ generation.

An example of the determination of the molecular ion of 23-

nor-gorgost-5-en-3beta-ol is given below. The voluminous output is included for illustrative purposes to see the operation of various parts of the program. In normal operation all but the conclusion of the program is omitted.

23-NOR-GORGOST-5-EN-3BETA-OL

[This spectrum includes some column bleed, e.g. m/e 405 at M+ - 7amu.]

SPECTRUM AFTER TRIMMING AND CLUSTERING —

M/E	INT	M/E	INT	M/E	INT	M/E	INT
44	25	55	1375	67	433	69	908
79	407	81	774	83	1165	91	367
93	459	95	660	105	531	107	601
109	360	119	384	121	341	123	207
131	240	133	554	135	251	143	215
145	465	147	356	158	231	159	480
161	351	171	179	173	242	175	191
185	152	187	180	189	166	197	116
199	172	203	87	211	217	213	418
217	231	228	181	229	495	231	273
239	135	243	210	246	137	253	247
255	395	267	147	271	667	272	672
273	280	281	1258	296	306	299	341
300	226	314	1877	328	345	351	92
352	40	369	64	370	34	379	114
380	38	393	311	394	166	397	104
405	393	412	503				

CANDIDATE MOLWTS & SCORES

```
      .
      .
      .
 407  0.49  0.00
 408  0.51  0.00
 409  0.45  0.00
 410  0.40  0.00
 411  0.43  0.00
 412  0.43  0.00        [correct M+ not particularly highly
 413  0.39  0.00        ranked.]
 414  0.33  0.00
 415  0.37  0.00
      .
      .
      .
```

CANDIDATE M+s AND BF/DBF RATINGS AFTER MOLTST [Now start to make use of chemical information, like do we expect even or odd parity M+, should the molecular ion be there, candidates at masses below highest mass observed are less likely etc]

```
      .
      .
```

```
      .
408   0.51  0.58
409   0.45  0.71
410   0.40  0.58
411   0.43  0.71
412   0.71  0.00        [Correct M+ is doing quite well, it
413   0.39  0.71        occurs, it is most intense in its
414   0.33  0.58        group etc.]
415   0.37  0.52
      .
      .
      .
CANDIDATE M+s AND BF/DBF RATINGS AFTER CHMFLT
      .
      .
      .
408   0.75  0.95
409   0.72  0.95
410   0.52  0.93
411   0.77  0.91
412   0.95  0.56        [although belief in 412 increased by some
413   0.54  0.96        good losses, we also have bad losses
414   0.33  0.98        (e.g. loss of 7 to 405) that increase
415   0.68  0.97        disbelief.]
      .
      .
      .
TOP RANKED MOLECULAR WEIGHTS AND THEIR "CF" SCORES
412    0.38
```

2.6.1 Summary of Results From New MOLION Program

Results from four experiments with the MOLION program are summarized in Table III below.

Experiment	Compounds processed.
A	sterols.
B	acid methyl esters.
C	acid TMS esters.
D	amino acid TAB esters.

The table distinguishes cases where the molecular ion was correctly identified by being the highest ranked candidate (usually with a CF score considerably greater than any alternative candidate) and cases where the correct molecular ion was merely listed in the top ten candidates (in these cases all the top ten candidates having approximately equal CF scores).

	COMPOUND CLASS			
	A	B	C	D
M+ present & identified :	44	8	14	4
M+ present & listed :	7	5	1	-
M+ present and not identified:	1 (a)	4 (b)	2 (c)	-
M+ absent but identified :	-	6	18	7
M+ absent but listed :	-	1	-	8
M+ absent and not identified:	-	-	2 (c)	4 (b,d)

Notes on errors:

(a) errors due to ions recorded at masses considerably above M+
(b) errors due to impurities.
(c) errors due to simple parity tests failing to detect presence of nitrogen.
(d) errors due to mass differences of more than 115amu between the highest mass ion recorded and the true molecular weight.

Table III. MOLION Results with Four Classes of Compounds.

2.6.2 Proposed Additional Work on the MOLION Program

The MOLION program is currently being implemented on the PDP11/45 based computerized GC/MS systems in the Departments of Chemistry and Genetics. The program will be evaluated through tests on the analysis of urine samples and other body fluids. Further developments of the system will be made in accord with the results of these tests.

2.7 Congen Improvements

During the past year many improvements have been made in the version of CONGEN available for outside use. These improvements allow the user more flexibility and range in the use of existing commands. Further, some new commands have been created which increase the power and utility of CONGEN. The program has become easier to use and more robust. Finally in almost every subsection of the program the user can inspect the computation as it proceeds. This means that fewer long, wasteful computations will be performed.

2.7.1 Error Detection in Substructure Definition

We now differentiate between two types of substructures called patterns and superatoms. This simplifies the chemist's

interaction with the program in that it makes explicit the two different ways in which substructural information is used in the current version of CONGEN. Further, this distinction helped us write very complete error detecting routines in a relatively small number of lines of code.

When the chemist indicates that he(she) is finished defining a substructure by typing "done" in editstruc, we check the substructure for errors. If we find errors we ask the chemist whether or not he chooses to fix them. However, if the chemist tries to enter this substructure on the composition list or on the constraints list without fixing it we indicate that the errors have not been fixed. Further we ask him if he would like to fix them. If he says yes, we put him inside Editstruc. In Edit-struc there is a new command called ERRORS which will print out all of the errors made in definition.

If the chemist does not choose to fix the errors, we warn very clearly that the results will be unpredictable or erroneous. We allow the chemist to go ahead on the philosophy that he may have a perfectly good reason for doing what seems to us to be nonsense.

Some examples of the types of errors detected are:

a) x names and polynames in superatoms;

b) undefined atom names in patterns or superatoms;

c) free valences on patterns;

d) no free valences on superatoms;

e) an atom with too many attached bonds or a conflict between the hydrogen range specified and the number of free valences specified;

f) the lack of exactly one tag in a proton constraint; and

g) problems connected with use of multiple link nodes.

With link nodes we detect when one link node is illegally bonded to another link node, when a link node wrongly has more than two neighbors, when a link node is monovalently bonded, and finally when a link node has a tag. Much remains to be done in extending and integrating the link node concept through out the rest of congen. There also needs to be further error checking to warn users who change a superatom and then call the generate routines with out redefining composition. We have concentrated our efforts on mistakes which we have observed frequently when other chemists use CONGEN. Moreover, this error checking will serve to reduce substantially the errors made by chemists using the program.

2.7.2 Depth-First Imbedder

The IMBEDDER program was completely rewritten. Four major improvements were implemented and the efficiency of almost all of the different subsections was improved. First, the method of computation was changed from breadth first (all structures delivered at the same time) to depth first (the structures delivered one at a time as they are created). The chemist can now check the computation as it proceeds by using the cntrl-S and cntrl-I features. Use of the cntrl-S feature often will allow the chemist to see that a certain computation is much larger than he anticipated and to stop it before computer time is wasted. Use of cntrl-I allows the chemist to see if the imbedding is proceeding in producing the kind of structures he anticipated. If not, the computation can be stoped and the problem redefined with only minimal loss of human and computer time. Previously the chemist would have had to have waited excessive amounts of time before seeing any results only to find, for example, that another constraint should have been used or, for another example, that more pruning should have been done before imbedding. This new improvement will result in the saving of large amounts of computer time.

Second, all the constraints testing during imbedding is now done in the SAIL portion of CONGEN and structures violating the constraints are not returned. Previously all structures were returned to the LISP portion of CONGEN before any constraints checking and subsequent pruning were done. This new approach represents a real gain in efficiency because these programs run much faster in SAIL than they do in LISP.

Third, the canonicalization routines were rewritten. When they were first written there was the prospect that our system might be interfaced with Chemical Abstracts. With this prospect in mind, the canonicalization was done using a modified form of Morgan's algorithm which was easy for people to understand and closely related to the canonicalization algorithms used at Chemical Abstracts. Recently, the procedures were redesigned with efficiency as the only criterion. New algorithms were found and the process of assigning a canonical number to a structure is now much less costly in terms of time. Further, two structures which are aromatically equivalent are given the same key (related to its canonical number). Previously they were given different keys and special methods in Lisp were needed to insure their equality. Since many different parts of CONGEN use the canonicalizer this resulted in a gain in efficiency for all of them.

Fourth, a change was made so that any number of superatoms can be imbedded at one time. This means when large numbers of superatoms need to be imbedded the chemist can in one set of commands perform the entire task, rather than the more time consuming approach of one-at-a-time. However, the chemist can still choose for reasons of efficiency to imbed a single

superatom when special tests on the environment of that superatom are required. This also provides the opportunity for large, multiple imbeddings to be done in batch mode. (The batch command was rewritten so that it would accept multiple superatoms.) The large batch job is then run after midnight when the load average is low to increase the amount of computer time for other uses.

2.7.3 EDITSTRUC Changes

The RENUMBER command was added to give the chemist flexibility in choosing schemes of numbering the atoms in the structure. There have been internal changes made to the editstruc commands BRANCH, LINK, CHAIN, and DELATS. All involve the method of numbering atoms. It is now possible to create a substructure which has "gaps"(missing numbers) in its numbering to atoms. These changes necessitated some further changes in the routines which prepare and send structures to the IMBEDDER in CONGEN.

2.7.4 BATCH

The BATCH command was rewritten to take advantage of the fact that the new lower fork programs can accept any number of Superatoms to be imbedded. As the system load continues to increase BATCH will become a more attractive option.

2.7.5 RESTORE

The RESTORE command was changed so that files written by REACT can be restored as well as files written by CONGEN. REACT users make use of the new EXAMINE subcommand (discussed elsewhere in this report) as well as the mass spectral ranking program MSRANK. Therefore it is very natural for a REACT user to save results using the SAVE subcommand in REACT and later to restore these structures in CONGEN in order to examine or rank them. The reason for the incompatibility between REACT format and CONGEN format is that REACT save files contain often many different structure lists whereas CONGEN save files contain only one. Thus the RESTORE command in CONGEN must ask the REACT user which structure list to restore.

2.7.6 TREEGEN

The TREEGEN command was rewritten to ensure that no duplicates are produced. Duplicates arise if there is a superatom on the composition list and further if from the remaining atoms a copy of that superatom or a portion of that superatom can be constructed. Another way of saying this is duplicates arise if there is a pattern in the superatom and that pattern can be built

from the remaining atoms in the composition list. The new routines check for a superatom on the composition list and if there is one the canonicalization routines are called for each structure. This key is used to determine if the structure has already been added to the structure list. Several functions had to be rewritten to insure that this was done efficiently.

2.7.7 SURVEY AND EXAMINE

The command SURVEY was added to the experimental CONGEN and routines were written which allowed the chemist to look over the structure list for certain functional groups and other features defined by the. SURVEY has since been incorporated into a command called EXAMINE (see section 2.3).

2.7.8 DRAW

The draw command was extensively rewritten so that it would be more flexible. The user can either draw the whole structure list or give the command an argument which gives the range of structures to be drawn. The range given may or may not use the number associated with the structure as a key. The user can also give ranges such as 1-3 which means draw the first through the third structures on the list.

2.8 CONGEN Efficiency

We have a continuing long term commitment to improve the efficiency and the reliability of CONGEN. Algorithms under development are written quite differently from the way they should be rewritten to execute efficiently and reliably. For example, it is very natural to use free variables in the development of new code, but eventually when the function is assumed or proven to be working correctly these free variables should be eliminated to buy efficiency and modularity. LISP's inherent inefficiencies can often be circumvented by careful reprogramming. The project of block compiling CONGEN described below is a major step forward in providing an efficient and robust base for future development.

At the outset we had hoped that block compiling would speed up constrained structure generation by a factor of at least four. We ended up after a great deal of fine tuning with a speed-up factor of about two, but much higher factors in other parts of the code.

At the beginning of this project we used the new LISP subsystem Masterscope to analyze each of CONGEN's twenty one files. For each file we prepared a database for that file which

contained information about its functions and their variables. These databases are important for maintenance and ease of learning CONGEN as well as for their short term purpose of facilitating block compilation. For any function on a file which we have analyzed we can ask the database which variables that function uses freely and which other functions it calls. When all the files have been analyzed we can find out which functions call a particular function.

Further we developed automatic batch programs to make and test new versions of CONGEN and to update the supporting databases. These programs can be run at night when the system is lightly loaded. This helps spread out the load on our heavily used system and leads to improved quality control in the version we supply to users since our testing can afford to be much more extensive and thorough.

Now that we have finished the work of block compiling CONGEN much testing remains to be done to insure that no new errors have been introduced. Once we are satisfied that the block compiled version is robust and reliable it will replace the version of CONGEN which we distribute to outside users. When we have reached this stage the new block compiled version will be used for development as well. Blocks which are under rapid development can be substituted into the block compiled version in their interpreted or normally compiled form. New blocks can be added without disruption of the existing program organization. Hence in gaining efficiency we have not lost flexibly.

The work on block compiling was done with the help and direction of Larry Masinter, a former member of the DENDRAL project, now with Xerox Palo Alto Research Center.

2.9 CONGEN Reprogramming

We have been investigating the reprogramming of CONGEN into an Algol-like language. The goals of reprogramming are threefold: first, to unify the program into a single language which can be used on a variety of computer systems; second, to begin to compact the program into a manageable, cost-effective size for current time-sharing systems; and third, to improve typical runtimes for CONGEN so that it becomes a more attractive means for scientists to solve structure elucidation problems. A version of CONGEN which fulfills these goals would be useful on a variety of computer systems and could be exported to many different chemical and biochemical laboratories.

2.9.1 Unification Into a Single Language

CONGEN is currently coded in three different programming

languages. The constrained cyclic structure generator, which is the basic algorithm responsible for the generation of structures, the entire user interface and a number of control routines necessary to support communication between the three languages are all coded in Interlisp. Interlisp is a list-processing language, and the sections of CONGEN written in this language are heavily oriented toward the use of lists as data structures. The part of the program which deals with the drawing of structures, either on a teletype or on a graphics terminal, is coded in FORTRAN. The remaining parts of CONGEN, including those parts of the program responsible for obtaining final structures from intermediate representations and various routines to support these functions, are all written in SAIL, an ALGOL-60 variant designed for the PDP-10 computer.

Although it would be possible to emulate Interlisp's list processing facilities using, for example, the REFERENCE and RECORD capabilities of SAIL, initial timing tests have shown that no significant increase in speed could be obtained by such a move. It is felt that this is a reflection of the fact that Interlisp is tuned for list processing, and that SAIL merely provides list processing as a "add-on" feature to ALGOL.

We believe that it will be possible to unify CONGEN into one ALGOL-like language by utilizing data structures more suitable in such a language. It would be desirable to use a language which provides as little overhead as possible to the size of a running program. Although the mathematics of the algorithms in CONGEN is quite complex, the algorithms themselves make no complex demands on a programming language. Our initial choice of language which we are exploring as a vehicle for reprogramming, BCPL, is discussed in more detail in a subsequent section.

2.9.2 Compaction Into a Manageable Size.

The Sumex computer system facilitates rapid development of complex programs: a virtual memory, paging environment with a full 256K core image available to each of CONGEN's different language segments. We estimate that a fairly direct translation (i.e., with a minimum of redesign of the algorithms beyond replacement of lists with arrays) would likely result in a program requiring approximately 300K words of memory in which to run. This is too large, even with extensive use of overlays.

With a certain amount of theoretical work, we can develop an algorithm related to one discussed by Sasaki. We have made significant progress on this problem (see below). The algorithm will need to be mathematically proven and made suitable to handle the majority of problems with which CONGEN is typically presented.

Using an adaptation of the Sasaki algorithm, and redesigning the current SAIL and FORTRAN portions of CONGEN, we expect that an overlaid version of the new CONGEN would need on the order of 20-30K words of core to run on a PDP-10.

Our preliminary estimates are of course subject to uncertainty. They presume an external device (random access disk) for storage of structures. Large problems (1-2000 structures) would require temporary files totaling about 100 pages (512/PDP-10 words per page). The whole package in one core image would require 51K words. Any mechanism for overlays would make the largest core image required about 21K.

2.9.3 Improvement in Runtime

In addition to decreasing the program size, it is also necessary to minimize execution time. An experimental version of the algorithm mentioned above has been written in BCPL. We have obtained initial timing information for this algorithm and compared results with the current CONGEN. The test cases used were representative of the types of problems with which the program will be confronted. The new generating algorithm is designed to replace the Interlisp structure generator. For the typical problem involving real structures, the generation problem deals with Superatoms. Thus, the empirical formulas in Table IV represent a whole class of problems. For example, the time to perform $C_2H_2N_2O_2$ represents the time for not only isomers of this formula, but also the time for any problem with two tetravalent, two trivalent and two bivalent Superatoms together with two hydrogen atoms.

Table IV. Preliminary Timing Estimates[a] for Isomer Generation

Empirical formula	Number of Isomers	Time in seconds for generation Interlisp CONGEN	BCPL algorithm
$C_2H_2N_2O_2$	506	233	10.0
C_6H_6	217	113	9.5
N_{10}	91	52	70.5

[a] The times were obtained by running the respective test cases on lightly loaded DEC KI-10's. The values obtained for the CONGEN in Interlisp were obtained on a system operating under Tenex. The values obtained for the new BCPL algorithm were obtained on a system operating under DEC's TOPS-10 operating system. The values presented are the average of three runs, but must be viewed only as approximate because of variations in system overhead, e.g., paging, expected during normal use.

The timing values in Table IV are what we expected given

our knowledge of the two algorithms. It is characteristic of a Sasaki-like algorithm that as the number of atoms (or Superatoms in our implementation) of the same type increases, execution time increases exponentially. The considerations of symmetry built into the Interlisp version treat such problems more efficiently, so the increase in execution time is somewhat less than exponential. There is a point where the efficiency of both algorithms would be the same. This is illustrated by the data in Table IV. With diverse atom types the BCPL algorithm is 23 times faster for the example case $C_2H_2N_2O_2$. With six tetravalent atoms or Superatoms of the same type, the BCPL version is only about ten times faster. For N_{10}, where all atoms or Superatoms are of the same type and the same degree (same number of non-hydrogen neighbors) the Interlisp version is faster. It is characteristic of most problems with which CONGEN is confronted that there is a diversity of types of Superatoms. In our opinion, the factor of 10-25 in increased speed for such problems justifies using the new algorithm.

Work is also currently underway on a separate imbedding package, similar to the package in the current CONGEN, but restructured for efficiency. This, together with the structure drawing program will place all of CONGEN in a common language. We can report that as of Feb. 5, 1978, the first version of the imbedding algorithm in BCPL is running. Work is now under way to compare results with the production version of CONGEN to ensure the accuracy and reliability of the new imbedder.

2.9.4 Choice of Language for Reprogramming

Recursion seems essential to the clear phrasing of most of the algorithms, both in future development work and in the initial reprogramming effort. Since provision for recursion in FORTRAN is usually by an add-on package or other such assembly of special routines, and since this facility is not available on the machine on which CONGEN is being developed, FORTRAN is not viewed as a likely candidate as a language choice. In a similar vein, many ALGOL implementations are quite inefficient in handling recursion. In many of these ALGOL implementations, due to the fact that any function is allowed to be recursive, one must pay the price of recursion even for non-recursive portions of the program. Two ALGOL based languages which are exceptions to this generalized method for handling recursive routines are SAIL and BCPL. SAIL requires one to explicitly declare a procedure as recursive, and then to use a different calling technique than is used for non-recursive procedures. SAIL also provides more explicit control over allocation of recursive variables.

BCPL is a "mini-ALGOL" with the same block structure and looping statements as ALGOL but with much more limited semantics. BCPL compilers are generally small and simple, and are available

on a variety of machines. We feel that there are three advantages to BCPL. First, because its structure is simple one is to some extent insulated from collapse of or changes in the supported compilers. It would probably not take over a month for someone experienced in compiler implementation to construct a basic BCPL compiler from scratch. Thus, if programs are restricted to a fairly pure form of the language, even a total removal of support from the language by outside agencies would not be fatal to those programs. Secondly, the BCPL run-time system is generally quite small, and adds little overhead to the size of a running program (on the order of a few thousand words of core storage, compared to a few tens of thousand words of core storage for SAIL). Lastly, because BCPL is closely related to both SAIL and ALGOL, it would be fairly simple and largely mechanical to convert the BCPL code into its SAIL or ALGOL equivalent, if such a translation proved necessary or desirable.

Largely as a result of the effort required to analyze adequately the questions posed by the conversion study, work has already begun on the initial stages of CONGEN reorganization and translation. In an effort to study the the effect of BCPL, the target language, on program efficiency, as well as to study the speed of the generation algorithm described above a modified implementation of the algorithm was coded into BCPL. This code formed the basis for the estimates of speed and size improvements described previously.

2.9.5 A Version of CONGEN for the Chemical Information System

We have recently been discussing the prospects of a version of CONGEN available to the public on a fee-for-service basis as part of the NIH/EPA supported Chemical Information System (CIS). Discussions with Dr. Steve Heller, EPA, and Dr. William Milne, NIH, several months ago resulted in a contract with Stanford University to investigate the feasibility of translation of CONGEN into a language which could eventually be supported by the CIS. The outcome of this study was that such a translation was possible given the limited goals of the task. The previous section on reprogramming languages and progress was taken in part from the results of that study. We are currently drafting a detailed contract proposal to carry out the translation.

The limited goals include providing CIS with a version of the CONGEN program including some (but not all) of its current capabilities. This version is to be written in an Algol-like language and must run on the Division of Computer Resources and Technology's (DCRT's) DEC PDP-10 at NIH.

It is important to contrast these limited goals with the more ambitious objectives of reprogramming as discussed in the previous Section: 1) The translation for the CIS will produce a

version which will run only on the DCRT/NIH system. It will presumably also run on similar PDP-10's using the TOPS-10 operating system. This, however, represents only a very small subset of computer systems available to the chemical community. Thus the CIS version will not meet our expressed objective of a version of CONGEN which is exportable to a large number of research groups; 2) the translation for the CIS will utilize the Basic Combined Programming Language (BCPL). This is an Algol-like language which will suffice for the CIS version. It is not clear that this language is optimum for a program which is to be widely distributed. The development of a more machine-independent language, such as MAINSAIL at SUMEX, may provide a much better vehicle for wide distribution, in which case our efforts under the current DENDRAL grant would be directed toward that language; and 3) the contract with CIS will have very limited provision for introducing new improvements in CONGEN and related programs (see Sections 2.2 and 2.4) to the DCRT/NIH version of CONGEN in BCPL. It is obviously essential to provide the chemical community with the most up-to-date version of CONGEN and related programs. Work under the present grant is directed at this broader goal.

At the same time there are obvious similarities to the CONGEN translation effort supported by CIS and the NIH under the current DENDRAL grant. We would be foolhardy to view these efforts as mutually exclusive. The major similarity is that the choice of language is subject to similar restrictions, meaning that some ALGOL-like language would be used for the exportable version for wider distribution. We feel that a translation of parts of the program, for example from BCPL into MAINSAIL, is a relatively simple task given the similarities of the languages. The translation efforts would, we feel, be synergistic, providing more rapid access of the chemical community to CONGEN and other programs.

3 THEORY FORMATION PROGRAMS - Meta-DENDRAL

3.1 Incremental Learning

In order to allow applying the Meta-DENDRAL program[3] to a wider range of chemically interesting problems, we have begun to remove one of the most important current program limitations - its inability to add piecewise to what it has learned. Meta-DENDRAL must currently process all training data at once, producing a set of rules which cover that data. The amount of training data which the program may examine when forming rules is therefore limited by available computer memory. We aim to give the program the ability to learn incrementally resulting in the following benefits:

1) The chemist will be able to generate rules from one set of training data, examine the rules, and if necessary obtain additional data for modifying and adding to the rule set. By examining the partial results produced at each step, the chemist may determine which additional training molecules are most appropriate for the next learning cycle. We expect the program to aid in making this decision by suggesting new training molecules whose spectra will resolve among rules which represent alternate plausible explanations of the observed data.

2) Since the amount of training data processed strongly influences the reliability of the learned rules, training on arbitrarily large data sets will allow Meta-DENDRAL to form more accurate rule sets than currently feasible.

3.1.1 The Approach

The proposed approach to incremental learning involves hypothesizing a set of rules on the basis of existing training data, then updating the rule set when new training data is provided. When existing rules apply incorrectly to new training molecules, these rules are modified. New rules are added to the rule set by applying the current one-pass Meta-DENDRAL program to the portion of the new training data which cannot be explained by existing rules. The figure below summarizes the proposed approach.

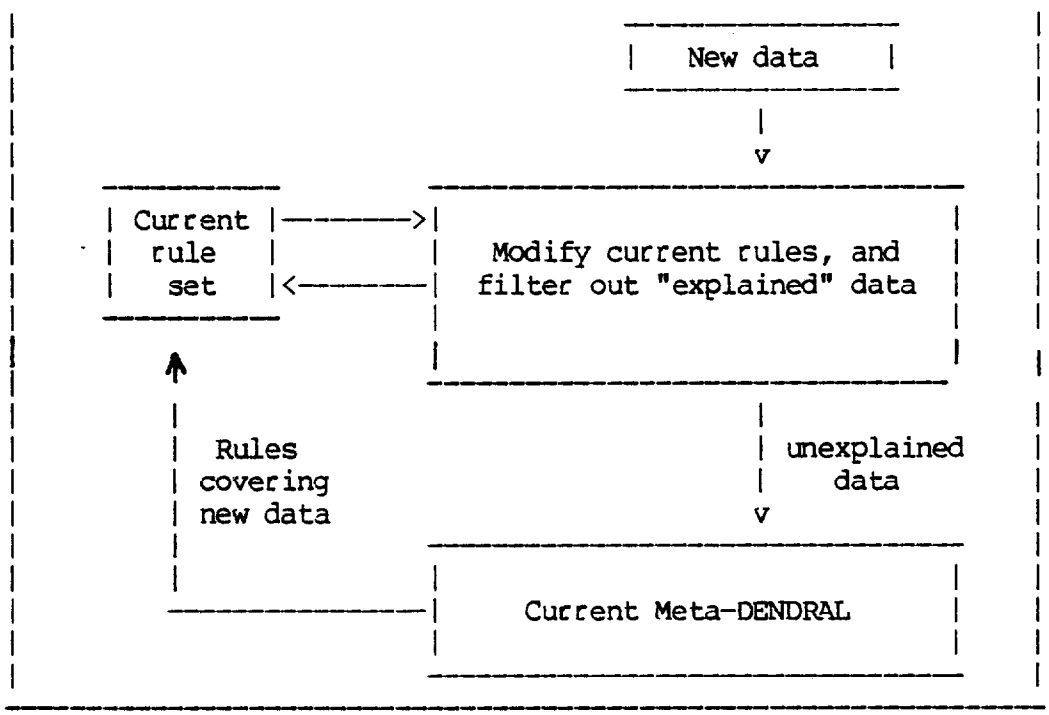


Figure 8. Approach to Incremental Learning

3.1.2 Modifying Existing Rules

Rules must be modified so that they become consistent with the new data while remaining consistent with previous data as well. In short, the method involves storing along with each rule a summary of alternate acceptable versions of the rule (those with the same evidential support in the observed training data). The summary of all acceptable versions of a given rule, refered to as the version space[12] of the rule, is useful for a number of tasks associated with rule learning, including incremental learning.

Version spaces provide an explicit representation of the space of all alternate versions of a given rule - i.e. those which cannot be disambiguated by the currently observed training data. As such, version spaces will allow Meta-DENDRAL to reason more thoroughly with the choice among alternate rule versions. Some expected benefits and uses of this increased ability are listed below.

Modifying current rules using new training data. Since version spaces provide a summary of all alternate versions of a given rule which are consistent with previous evidence associated with the rule, they delimit the range of allowed future

modifications to the rule. Thus, those modifications to the rule which are consistent with past data are exactly those which yield versions of the rule contained in the version space. At any point, the version space contains all rule versions which are consistent with a given set of evidence. The "best" rule version can then be chosen (e.g.,. on the basis of simplicity or chemical plausibility) from the entire space of rule versions consistent with the data.

More complete exploration of alternate rules. The current RULEMOD portion of Meta-DENDRAL tries to make rules more general or more specific in order to improve their performance on the training data. RULEMOD tries many such ways of modifying rules, but it cannot afford to try all ways. This portion of RULEMOD will be replaced by a new routine which will use version spaces to explore all ways of generalizing or further specifying rules in order to eliminate negative evidence or add additional positive evidence.

Intelligent selection of training instances. Since version spaces represent the range of plausible alternate versions of a rule, they contain the information needed to select new training instances designed to discriminate among competing rule versions. For instance, by examining a given version space the program ought to be able to suggest a set of compounds whose spectra would allow ruling out many of the plausible rule versions while strengthening the evidence associated with other versions.

3.1.3 Version Spaces

This section presents a sample version space generated by Meta-DENDRAL, and discusses how version spaces may be updated to take into account new training data. Notice that the program is dealing not with a single rule which will be later modified, but with the space of all plausible rule versions. The algorithm for updating version spaces using new training data is a **candidate elimination** algorithm: candidate rule versions are eliminated from the version space as they are found to perform incorrectly on new training data. The candidate elimination algorithm is assured of finding all rule versions consistent with a given set of positive and negative training instances. This is accomplished without backtracking and independent of the order of presentation of the training instances.

3.1.3.1 Definition and Representation

The key to an efficient representation of version spaces lies in observing that a general-to-specific ordering is defined on the space of rule subgraphs. The version space may be represented in terms of its maximal and minimal elements according to this ordering.

To see exactly how the general-to-specific ordering comes about, consider an example. Suppose that R1 and R2 are two rules which predict the same action. Then R1 is said to be more specific (or, equivalently, less general) than R2 if and only if it will apply to a proper subset of the instances in which R2 will apply. This definition is simply a formalization of the intuitive ideas of "more specific" and "less general".

The general-to-specific ordering will in general be a partial ordering; that is, given any two rules we cannot always say that one is more general than the other. Therefore, when all elements of the version space are ordered according to generality, there may be several maximally general and maximally specific versions.

Version spaces can be represented by these sets of maximally general versions, MGV, and maximally specific versions, MSV. Given such a representation it is quite easy to determine whether a given rule belongs to a given version space. A rule statement belongs to the version space of a given set of positive training instances and negative training instances if and only if it is (1) less general than or equal to one of the maximally general versions, and (2) less specific than or equal to one of the maximally specific versions. Condition (1) assures that the rule cannot match any training instance in I-, while condition (2) assures that it will match every training instance in I+. Since the sets MGV and MSV are by definition complete, (1) and (2) will be necessary as well as sufficient conditions for membership of a rule statement in the version space.

3.1.3.2 Example From C13 NMR Rule Formation

Meta-DENDRAL has been used to determine rules associating substructures of molecules with data peaks in a carbon-13 nuclear magnetic resonance spectrum [11]. Figure 9 shows a version space represented by the program in terms of the sets of maximally specific rule versions (rule MSV1) and maximally general rule versions (rules MGV1 and MGV2). This version space contains all rules which predict a CMR shift of from 14.0 to 14.7 ppm. downfield from TMS and which are consistent with a set of paraffin and acyclic amine data presented to the program. The rule pattern which expresses the conditions for application of each rule is stated in the language of chemical subgraphs. Each node in the subgraph represents an atom in a molecular structure. Each subgraph node has the four attributes shown, with values constrained as shown in Figure 9.

Rule Subgraph		Constraints on Subgraph Node Attributes			
subgraph	node name	atom type	number of non-hydrogen neighbors	number of hydrogen neighbors	number of unsaturated electrons
MSV1:					
v-w-x-y-z	v	carbon	1	3	0
	w	carbon	2	2	0
	x	carbon	2	2	0
	y	carbon	2	2	0
	z	carbon	>=1	any	0
MGV1:					
v-w-x	v	carbon	1	any	any
	w	any	2	any	any
	x	any	>=1	2	any
MGV2:					
v-w-x	v	carbon	1	any	any
	w	any	2	any	any
	x	any	2	any	any

Figure 9. A Version Space Represented by It's Extremal Sets

Notes to Figure 9:

MSV1 is the maximally specific rule version.

MGV1 and MGV2 are maximally general rule versions.

Only the rule patterns (left hand sides) are shown above.

All rules shown predict the same action: the appearance of a peak associated with atom "v" in the range 14.0 to 14.7 ppm. downfield from TMS.

The version space represented in Figure 9 above contains several hundred rule versions: the three versions shown plus all versions between these in the general-to-specific ordering. However, it can be represented simply by the two maximally general versions, MGV1 and MGV2, and the single maximally specific version, MSV1. The single most specific version contains every node and node attribute constraint consistent with all positive training instances. In this program the classes of positive and negative training instances are sets of molecules for which the indicated spectral peak does and does not appear. Thus, any rule version more specific specific than MSV1 cannot match every positive instance. Two general versions are required in this case since neither is "above" the other in the general-to-specific partial ordering. Any rule more general than either MGV1 or MGV2 will match some negative instance. Furthermore, any rule which is between these general and specific boundaries of the version space will match all current positive instances (by virtue of being more general than MSV1), and will match no current negative instances (by virtue of being more specific than MGV1 or MGV2).

3.1.3.3 Version Spaces and Rule Learning

Rather than select a single best rule version, the candidate elimination algorithm represents the space of all plausible rule versions, eliminating from consideration only those versions found to conflict with observed training instances. Thus, the candidate elimination approach separates the deductive step of determining which rule versions are plausible, from the inductive step of selecting a current-best-hypothesis. The algorithm is assured of finding all correct versions of the rule after all training data has been presented without the need to backtrack to reconsider previous training data or decisions.

In this example, RULEGEN was used to generate a set of plausible rules characterizing the CMR spectra of a set of

training molecules. For each rule, the associated evidence was given to a the candidate elimination routine which formed the version space for this evidence set. Subsequent data may be analyzed to modify the version space in a manner guaranteed to be consistent with the original data.

The candidate elimination algorithm operates on the maximally general and maximally specific sets representing the version space. The set of maximally general rule versions (MGV) is initialized to a single rule consisting of the most general possible rule subgraph (a single atom graph with no constrained node attributes), and the predicted shift range determined by RULEGEN. The set of maximally specific versions (MSV) is initialized to a rule which contains as its subgraph the entire molecule associated with the first observed positive instance. The initial version space represented by these extremal sets therefore contains all rules which match the first positive training instance (the most general possible rule, the very specific rule, and all intermediate rules).

The training instances are then considered one at a time. Each training instance is used to eliminate from the version space those rule versions which conflict with that instance. This is always accomplished by shifting the maximally specific and maximally general boundaries of the version space toward each other as shown in Figure 10.

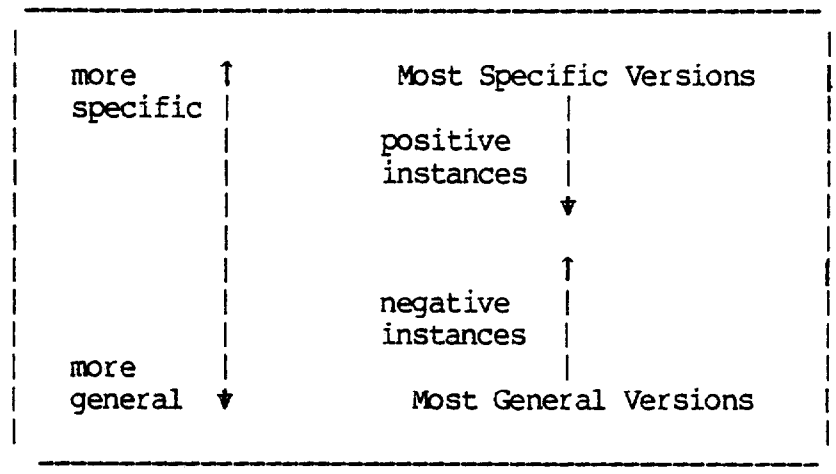


Figure 10. Effect of Positive and Negative Training Instances on Version Space Boundaries

Positive training instances force elements of MSV to become more general, whereas negative training instances force elements of MGV to become more specific. The maximally specific set can,

of course, never be replaced by a more specific set (nor the maximally general set by a more general one) since by definition, any version outside the current version space boundaries is inconsistent with previous training data. The action taken by the candidate elimination algorithm in updating the extremal sets is given below.

For negative training instances, each element of MGV which matches the instance must be replaced by a set of minimally more specific versions which do not match the instance. These new versions are obtained by adding constraints taken from elements in MSV in order to ensure that they remain more general than some MSV, and thus remain consistent with previous positive instances. Furthermore, each element of MSV which matches the negative training instance must be eliminated from the set (since it is already maximally specific, it cannot be replaced by a more specific version).

For positive training instances, any elements from MSV which do not match the new instance are replaced by a set of minimally more general elements which do match the instance. In order to ensure that these more general versions do not match past negative training instances, any which are not more specific than at least one element of MGV are eliminated. Elements from MGV which do not match the positive instance are eliminated.

After processing each training instance, the new maximally general and maximally specific sets will bound the space of all rules consistent with the observed data.

3.1.4 Current Status and Future Work

The incremental learning ability for Meta-DENDRAL depicted above in Figure 8 is almost fully implemented, but as yet remains untested. Routines for defining and modifying rule version spaces are implemented, as well as the ability to filter out training data explained by a rule set. The major unimplemented portion of the incremental learning scheme is the process for merging new rules into the evolving rule set. The chief issue here is deciding when and how to chose among or merge new rules which are similar to existing rules. We expect to complete implementation and initial testing of the incremental learning ability during 1978.

Among issues associated with the version space approach which we expect to explore during the current grant period are the following:

1) Intelligent selection of new training data from examination of partial results.

2) Applying chemical plausibility

information to select a "best" rule version from among those contained in the version space.

3) The extension of current methods for dealing more completely with noisy and ambiguous training data.

4) The use of version spaces for merging similar rules.

3.2 New Capability To Emphasize Discriminatory Power

One important intended use of rules formed by Meta-DENDRAL is the prediction of mass spectra for use in structure elucidation: Predicted spectra for a set of candidate structures are compared by computer with the mass spectrum observed for an unknown compound, and on this basis the candidates are ranked according the likelihood of their identity with the unknown. The ability of rules, in this context, to differentiate correctly among candidate hypotheses is called their "discriminatory power." Since the selection criteria previously used by Meta-DENDRAL during the various stages of rule formation did not necessarily correlate with high discriminatory power, it was decided to provide the program with the option of directly emphasizing discriminatory power during rule formation, in order to maximize the usefulness of the resulting rules for purposes of structure elucidation.

This addition to Meta-DENDRAL has now been designed and implemented. The general method employed by the the new option is as follows. Observed mass spectra of the training molecules are analyzed prior to rule generation to determine how diagnostic the various observed peaks are, within the training set, of the molecules that show them. This information is then used during rule formation to compute a measure of discriminatory power for emerging rules. This measure is used, in combination with other criteria, to guide the search during rule generation, and to control the modification and selection of rules during the later phases of processing.

Preliminary testing of this new rule formation scheme on the monoketoandrostanes produced rules of considerably greater discriminatory power within that family than had been produced in earlier work with Meta-DENDRAL, even though the training set used was only half as large as that used earlier. This "discrimination option", now integrated with the new template-processing capability, is currently being further tested on a group of aromatic esters to determine whether the rules formed are consistent with what is known about the fragmentation modes of those molecules, and whether the rules have significant discriminatory power outside the training set used to form them.

3.3 Improved Ranking Capability

The program used within the Meta-DENDRAL framework to rank candidate structures has been improved in several ways.

A) The program now summarizes its own results and prints the summaries, thus eliminating much tedious manual analysis that previously was necessary. This makes possible a much more systematic and extensive investigation of scoring functions and their behavior than was previously possible.

B) A large number of new scoring functions have been made available, many of them specially designed for use with rules formed under the "discrimination option."

C) A new ranking method has been implemented as an option, with an eye toward improving the application of scoring functions in ranking. This new method eliminates duplicate explanations of peaks (which were previously permitted) in a principled way. The new method may be easier to justify theoretically, and yielded generally better ranking results than did the old method in tests performed with monoketoandrostanes. Further tests are planned with aromatic esters and marine sterols.

3.4 Data Selection Program

It is a commonplace of methodology that good inductive generalizations depend on variety in the data set. This is no less true in the context of rule formation by Meta-DENDRAL. Whether the goal is to discover rules of high generality or high discriminatory power, one's chances of achieving this goal [appear to] increase with increasing variety of training instances. This suggests that it would be useful to have a data selection program that would select the subset of the potential training molecules which has the greatest variety, in some appropriate and well-defined sense.

A preliminary version of such a program has been implemented, and experiments with it will soon be underway. The method employed has two steps: A.) Construction of an index of all the structurally different possible fragmentation environments permitted in the molecules of the set of potential training molecules (PT) by the "half order theory" of mass spectral fragmentation. B.) Construction of an n-sized subset of PT that contains nearly the largest number of different permitted fragmentation environments possible for a set of that size.

3.5 Feedback Loops

3.5.1 Filtering with Respect to Existing Rules

The RULEGEN program is capable of accepting previously defined rules as a means of filtering the evidence obtained from INTSUM before the evidence is used for rule formation. As well as providing a convenient and natural feedback mechanism for the program, this facility also allows rules obtained from other sources to be used to reduce the space which the program must examine to find rules for a given set of data. In this manner, the program is able to focus attention on evidence which is not already explained by any of the rules which it is given.

A problem with this approach arises from the fact that the spectral evidence may often be the result of more than one fragmentation. Yet the filtering mechanism assumes that any evidence which supports a rule is completely accounted for by that rule. Tests are in progress to determine the limitations of this approach.

3.6 Program Improvements

3.6.1 Defining Rules with EDITSTRUC

In addition to the programs which produce rules from the spectral data, other programs have been developed to allow a user to define a set of rules manually. Like the rules produced by RULEGEN and RULEMOD, these are rules of structure fragmentation which are expressed in terms of molecular subgraph descriptions. The programs for manual rule definition provide a simple yet useful language for the description of these rules. A principle part of this language is the EDITSTRUC language, developed for CONGEN. This allows us to take advantage of the advanced structure manipulation capabilities which are a part of the EDITSTRUC package.

The ability to create rules manually should be particularly useful in conjunction with the rule filtering mechanism of RULEGEN mentioned previously. This provides the chemist with a natural means of describing obvious rules which the program can eliminate from consideration before focusing on the remaining unexplained evidence.

3.6.2 Stability Rules in INTSUM and RULEGEN

The programs have been generalized to allow the analysis of the mass spectral data from the point of view of determining

rules about stable bonds, i.e., lack of fragmentation in a molecule as well as fragmentation. Just as peaks are evidence of fragmentation in a structure, absence of peaks is evidence that certain fragmentations have not occurred.

The programs are now capable of examining the original data from either point of view and proposing rules of behavior of the molecules from that point of view. Further work remains to be done to carry this generality through the processing performed in RULEMOD and then in conducting experiments to determine the usefulness of stability analysis.

3.6.3 Expanded Template Space

Originally, the subgraph descriptions in the rules produced by the RULEGEN program were restricted by requiring that the internal connection patterns of the subgraphs had to be completely specified. In other words, for each of the interior nodes in the subgraph, the complete set of neighbors had to be specified. This restriction excluded rule forms which seemed to be both plausible and desirable, so the program was changed to eliminate the restriction.

In terms of the mechanism used by the program to search the space, implementation of this change meant removing the restriction on the subgraph matching templates that the neighbors property be required at all but the outer levels of a template. This allows the program to find rules in which the internal connection patterns of the chemical subgraphs are only partially specified.

For example; it is now possible to express a rule such as "break any bond which is 2 bonds away from an oxygen atom". Such a rule could not be expressed previously without identifying whether the nodes between the oxygen atom and the break were secondary, tertiary, or quaternary.

3.6.4 Small LISP and Program Efficiency

Increased size and complexity of the Meta-DENDRAL software has resulted in increasing efforts aimed at making the programs more efficient and understandable. All the programs which are part of the meta-DENDRAL system are now capable of running in the environment of "small LISP". This makes considerably more memory space available to the chemist for the data structures, thus making possible the solution of significantly larger problems than were possible in the standard LISP environment.

3.6.5 Help Facilities

As the programs have increased in complexity and usefulness, we have had to face problems of documentation and explanation of the programs to its users. Text explanations of the various aspects of the programs must be provided, and kept up to date, to allow others to use the system. It is also important that the text descriptions of the programs be available to the programs themselves to be used during program execution to provide on-line guidance to the user concerning the use of the programs.

Text descriptions of the programs must be closely associated with the programs themselves to insure that program changes are reflected accurately in changes in the text which describes them. Yet text explanations must be incorporated into the programs so as not to take up space which should be available during program execution to be used for producing results. Attempt has been made to resolve these sometimes conflicting goals through the use of the comment facilities of LISP, and through the generation of programs and conventions for programming which allow program documentation and explanations to be incorporated into the programs as comments in the appropriate places. There are then programs which have access to this information to produce documents and on-line explanations about the programs.

4 COLLABORATIVE RESEARCH

4.1 CONGEN Users

Dr. Peter Gund of Merck, Sharpe and Dohme Laboratories contacted us for a current CONGEN manual and Guest login information. He now feels that he has analytical problems which would lend themselves well to checking with CONGEN.

Professor Richard E. Moore of the University of Hawaii visited Stanford and was provided with a CONGEN demonstration on a problem relating to his own marine sterol work. We discussed system access and Tymnet node availability with him. He plans to return in the near future with another problem, and then consider the possibility of requesting access.

Dr. Jean-Claude Braekman of the University of Brussels travels across Brussels to use a terminal at the offices of the Belgium Chemical Society, in order to access CONGEN on SUMEX. Dr. Braekman uses the mail facilities to remain in contact with Prof. Djerassi's research group.

Dr. Martin Huber, a postdoctoral fellow in Professor Wipke's SECS group has been starting work in an area which was related to the graph theoretic basis for CONGEN. In an effort to encourage cross-fertilization or ideas, we encouraged and arranged a meeting between him and several of the DENDRAL project members. The resulting discussion, at the least, provided Dr. Huber with suggestions and information for further study. Likewise, DENDRAL was able to obtain a better idea of similarities in research interests between the two groups. We are currently pursuing several problems in graph theory concerning analysis of molecular structures. These problems arose directly from this meeting and concurrent discussions with Prof. Wipke.

During the special symposium at the San Francisco ACS meeting in the fall of 1976 which Ms. Suzanne Johnson helped to organize and chair, members of the DENDRAL group provided on-line demonstrations of CONGEN during the "hands-on" session. At this time Professor Kurt Mislow of Princeton University expressed interest in using the program. Later, we provided him with Guest access information and answers to his questions concerning terminals and other useful programs available to chemists on various commercial networks. As a result of this effort, Professor Mislow has used CONGEN and has been considering its use as a teaching aid. He wrote us this past spring to enquire whether Guest access to CONGEN might be possible for his friend Professor Weiss, head of the Department of Chemistry at Northeastern University. We subsequently provided Professor Weiss with the information necessary to access CONGEN on a trial basis.

In November 1976, Dr. Stan Lang of Lederle Labs' Infectious Disease Research Section, requested access to CONGEN. After being providing with the appropriate information and initial help, he encouraged Dr. Leon Goldman to request access also, and to request information on obtaining a copy of the teletype DRAW program used to draw CONGEN structures on teletypes. A recent phone conversation with Dr. Babu Venkataraghavan, a new member of the research group at Lederle, indicated that the TTY DRAW program was being used quite successfully. Also interested in the possibility of support for graphics terminals, Dr. Venkataraghavan called to discuss the problem in terms of Omnigraph, which they already have on their PDP-10. We have exported a complete copy of all the DRAW program files, including ample data files, to Dr. Venkataraghavan and are currently in contact with him on implementation questions.

A further example of cooperation between DENDRAL and Professor Wipke's group concerns the sharing of graphics programs. DENDRAL obtained the Fortran sources for programs created by the SECS group to do molecular modelling and structure display on the DEC GT40. Wanting to interface these programs to CONGEN, but not wanting to limit CONGEN graphics to one terminal

type, DENDRAL personnel modified the program to use the Omnigraph graphics package available on SUMEX. Glenn Ouchi of the SECS project, has become familiar with the relationship of the graphics in CONGEN to the Modeller's graphics. SECS has become aware of the desirability of supporting additional terminal types for graphics output, and will be investigating Omnigraph applications to this area.

One of the students who used CONGEN in Prof. Djerassi's molecular structure elucidation course introduced the program to a graduate student of Professor E.J. Eisenbraun's (Oklahoma State University). Professor Eisenbraun is a well known marine natural products chemist. He has requested Guest access information, and appropriate materials were provided in spring of 1977. Professor Eisenbraun subsequently visited Stanford and got a personal demonstration of CONGEN.

We have been in contact with Dr. Karl Kuhlman, a chemist and PROPHET user at SRI International. We have arranged for a group of DENDRAL chemists to get together with the SRI group for exchange demonstrations: CONGEN for PROPHET, and discussion of similar problem areas with visiting PROPHET representatives.

Dr. David Pensak of Dupont in Wilmington, Delaware originally started out as a CONGEN Guest user. In return, he contributed a good deal of knowledge concerning evaluation and use of molecular modelling programs. At the current time he is beginning to a build a research group in computer applications in chemistry, and views SUMEX/DENDRAL somewhat as a resource from which to obtain knowledge of hardware, software and people.

Dr. Milton Levenberg of Abbott Laboratories first expressed interest in CONGEN at an ACS meeting two years ago. He was given an account and appropriate information at that time. He had used OMNIGRAPH to develop a program to display and plot mass spectra, which he gladly provided to us. That program now provides a means for chemists to obtain a plot of their spectra which have been obtained on mass spectrometers which are not yet equipped with automatic computer output.

When Kent Morrill was a graduate student in chemistry he developed an interest in CONGEN and various of the Meta-DENDRAL programs. When he left recently for a job with Tennessee Eastman, he requested Tymnet login information to take with him. As a result of his interest, Dr. Gary Santee of Eastman Kodak in Rochester requested information for Guest access to CONGEN. Kodak may also be in the process of forming a computer applications in chemistry group, and once again, we seem to be viewed as a potential information resource in this type of effort.

Dr. Gretchen Schwenzer was a postdoctoral fellow with DENDRAL. When she left Stanford for a job at Monsanto, it was

with the idea of taking part in helping to develop a computer applications in chemistry group. She too views SUMEX as an information and know-how resource. To that end, we have had several phone calls and terminal links from her concerning graphics, terminals, modelling programs and text editors. She is interested in obtaining several copies of documentation preparation programs either developed or supported at SUMEX.

Dr. Robert Shapiro of New York University came to visit Stanford in September of 1977 to learn to use CONGEN. He spent a week in residence to discuss structure elucidation problems relating to nucleic acids and their interactions with other substances. We are also pursuing ideas on the automated analysis of UV spectra of such compounds, based on empirical rules derived from study of known systems.

In November of 1976, Dr. Henry Stoklosa of Ciba-Geigy approached one of the members of the DENDRAL project for trial use of INTSUM. During a subsequent visit to Stanford, we introduced him to CONGEN and its use. We have been keeping him up to date on recent developments because he indicated that CONGEN is beginning to have more and more use to him in the analytical task of evaluating additive bonding in polymeric materials.

Dr. Geza Szonyi of Polaroid corporation was one of the original persons to enquire about SUMEX/CONGEN access as a result of the "invitations for use" which were included as a part of early journal articles. He has recently requested trial access to CONGEN. Phone conversations indicate that his group is evaluating computer systems which will offer them the greatest latitude in applying computers to their work in various fields of chemistry and related data management. Once again, DENDRAL is viewed as a potential knowledge source.

Drs. D. Williams and R. McGrew from the Midland, Michigan site of Dow Chemical came to visit Stanford and receive an introduction to CONGEN. They were given a CONGEN demonstration, and as a result, requested a copy of the teletype DRAW portion of the program, which we sent to them. This brings to five the number of sites which are now using the teletype DRAW program in some fashion. Also included are: Lederle Labs in New York, (Dr. Babu Venkataraghavan); Dept. of Computer Science at SUNY, (Dr. Dave Larson); Dept. of Chemistry, Arizona State Univ., (Prof. Morton Munk); Dept. of Chemistry, Miyagi Institute, (Prof. Hidetsugu Abe); and Cambridge University, (Neil Gray).

4.2 Marine Natural Products

4.2.1 Mass Spectral File Search System

An attempt was made to obtain mass spectra for all marine sterols reported in the literature (Appendix A). The old mass spectral files were scanned and pertinent sterol mass spectra were digitized (a file of non marine sterol mass spectra were also acquired from the older files as a supplement to the marine file) (see Appendix B. Marine sterol researchers were requested to send samples of specific sterols which they reported or sterol mixtures known to contain the requested sterol (see Appendix B. In a few cases sterols were isolated from crude extracts of organisms known to contain the sterols. The high resolution GC-MS spectra of the available sterols were recorded using a Hewlett Packard 7610A gas chromatograph equipped with a 10' X 2 mm "U" shaped column (3 per cent Poly S-179 on gas chrom Q or 3 per cent OV-17 on gas chrom Q (column temp. 260 degrees C) and interfaced with a Varian Mat 711 double focussing mass spectrometer (equipped with a Watson-Biemann dual stage separator, an all glass inlet system and a PDP-11/145 computer for data acquisition). High resolution spectra were recorded for subsequent fragmentation analysis by the application of date interpretation and summary programs, e.g.,. INTSUM, and to facilitate handling of the data for construction of the searchable files. Within the framework of the available data acquisition and reduction systems, the rapid analysis scheme has been tested, and the advantages and limitations are the subject of the following section.

The spectra of 52 marine sterols were compiled in a computer searchable format. The spectra, which are essential to have available for careful comparison following the search report, have been plotted, and the plausible or established interpretations of the higher molecular m/e peaks have been indicated on the spectra. Spectral interpretations have been coded in Fig. 8 in a series of 32 symbols which have been appropriately marked on the spectra of each sterol in Appendix C which is the file of marine sterol spectra constructed in our laboratory. Attached is a list of investigators who reported and received copies of this file. This summary of proposed fragmentation rules is acting as a preliminary guide in the INTSUM evaluation.

The SEARCH program was used to match every spectrum in the file (Appendix C) to every other spectrum to gain an indication of how all the spectra rank to one another in terms of the similarity index described previously (Table V). A rank of 999 indicates a positive identification; therefore, each spectrum when compared against itself results in a rank of 999. Ranking values below 500 indicate positive nonidentity and are not recorded. Ranking values approaching 750 indicate a possible match is not ranking higher due to variations in spectrometer operating conditions. Table V displays a number of interesting results. First, several separate sterols rank at the identity

rank, that is, they have mass spectra which are similar enough to be basically indistinguishable:

Sterols 15 and 18 Appendix A: this indicates that mass spectrometry cannot distinguish between slightly different side chain alkylation patterns in some cases. This agrees with the similar evaluations in the literature.

Sterols 68 and 71: this indicates that mass spectrometry cannot distinguish between side chain double bond geometrical isomers (E and Z) in this case.

Sterols 90 and 80: these are again sterols with slightly different patterns of side chain alkylation.

See pp. 88a-c for Table V.

Table V.

LIBRARY SEARCH REPORT FOR EXPERIMENT SEARCHING
52 SPECTRA IN MARINE AGAINST THEMSELVES

STEROL #	RANK	STEROLS MATCHED STEROL
1	999	87 999 0 274 ANDROST-5-EN-3BETA-OL
2	999	92 999 0 300 PREGNA-5,20-DIEN-3BETA-OL
3	999	99 999 0 300 PREGNA-5,17(20)Z-DIEN-3BETA-OL
4	999	55 999 0 302 PREG-5-EN-3BETA-OL
5	999	90 999 0 314 23,24-DINOR-CHOLA-5,20-DIEN-3BETA-OL
	547	52 0 0 412 24-ETHYLCHOLESTA-5,24(28)Z-DIEN-3BETA-OL
	554	41 999 0 426 (24Z)-24-PROPYLIDENECHOLESTA-5-EN-3BETA-
6	999	107 999 0 316 23,24-DINOR-CHOL-5-EN-3BETA-OL
7	999	101 999 0 318 5ALPHA-23,24-DINOR-CHOLAN-3BETA-OL
8	999	91 999 330 330 24-NOR-CHOL-5-EN-3BETA-OL
9	999	99 982 0 370 24-NOR-CHOLESTA-5,22E-DIEN-3BETA-OL
	508	60 999 0 370 5ALPHA-24-NOR-CHOLESTA-7,22E-DIEN-3BETA-
13	999	109 999 0 372 24-NOR-CHOLEST-5-EN-3BETA-OL
10	999	118 999 0 370 5ALPHA-24-NOR-CHOLESTA-7,22E-DIEN-3BETA-
	576	68 999 0 398 5ALPHA-24-METHYLCHOLESTA-7,22E-3BETA-OL
	555	55 984 0 370 24-NOR-CHOLESTA-5,22E-DIEN-3BETA-OL
	504	56 999 0 384 CHOLESTA-5,24-DIEN-3BETA-OL
14	999	124 999 0 382 CHOLESTA-5-EN-23-YN-3BETA-OL
15	999	114 999 0 384 (24S)-27-NOR-24-METHYLCHOLESTA-5,22E-DI
	800	76 0 0 384 CHOLESTA-5,22E-DIEN-3BETA-OL
	612	68 999 0 384 CHOLESTA-5,24-DIEN-3BETA-OL
	518	55 999 0 398 24-METHYLCHOLESTA-5,22E-DIEN-3BETA-OL
16	999	89 999 0 384 (24S)-5ALPHA-27-NOR-CHOLESTA-7,22E-DIEN-
17	999	113 0 0 386 (24S)-5ALPHA-27-NOR-24-METHYLCHOLEST-22E
18	999	95 0 0 384 CHOLESTA-5,22E-DIEN-3BETA-OL
	745	85 999 0 384 (24S)-27-NOR-24-METHYLCHOLESTA-5,22E-DIE
	594	66 999 0 384 CHOLESTA-5,24-DIEN-3BETA-OL
	547	58 999 0 398 24-METHYLCHOLESTA-5,22E-DIEN-3BETA-OL
23	999	111 999 0 384 CHOLESTA-5,24-DIEN-3BETA-OL
	631	72 999 0 384 (24S)-27-NOR-24-METHYLCHOLESTA-5,22E-DIE
	642	61 999 0 384 CHOLESTA-5,22E-DIEN-3BETA-OL
	581	68 999 0 426 GORGOST-5-EN-3BETA-OL
25	999	96 999 0 386 CHOLEST-5-EN-3BETA-OL
	647	55 999 0 386 5ALPHA-CHOLEST-7-EN-3BETA-OL
26	999	85 999 0 386 5ALPHA-CHOLEST-7-EN-3BETA-OL
	593	57 999 0 386 CHOLEST-5-EN-3BETA-OL
	581	58 999 0 400 5ALPHA-24-METHYLCHOLEST-7-EN-3BETA-OL
	500	37 999 0 426 (24Z)-24-PROPYLIDENECHOLESTA-5-EN-3BETA-
27	999	104 999 0 386 5ALPHA-CHOLESTAN-3BETA-OL
	629	68 999 3 388 5BETA-CHOLESTAN-3BETA-OL

STEROL #	RANK	STEROLS MATCHED			STEROL
28	999	105	999	0 384	CHOLESTA-5,7-DIEN-3BETA-OL
30	999	82	999	0 372	19-NOR-CHOLEST-5-EN-3BETA-OL
32	999	108	999	0 388	5BETA-CHOLESTAN-3BETA-OL
	673	70	999	0 388	5ALPHA-CHOLESTAN-3BETA-OL
38	999	106	999	0 398	24-METHYLCHOLESTA-5,22E-DIEN-3BETA-OL
	669	79	999	0 398	5ALPHA-24-METHYLCHOLESTA-7,22E-3BETA-OL
	614	70	0	0 398	(24S)-24-METHYLCHOLESTA-5,25-DIEN-3BETA-
	581	68	999	0 426	GORGOST-5-EN-3BETA-OL
	556	59	999	0 398	24-METHYLCHOLESTA-5,24(28)-DIEN-3BETA-OL
41	999	104	999	0 396	24-METHYLCHOLESTA-5,7,22E-TRIEN-3BETA-OL
	558	48	0	0 410	24-ETHYLCHOLESTA-5,7,22E-TRIEN-3BETA-OL
43	999	114	0	0 398	(24S)-24-METHYLCHOLESTA-5,25-DIEN-3BETA-
	660	70	999	0 398	24-METHYLCHOLESTA-5,22E-DIEN-3BETA-OL
	635	75	999	0 398	5ALPHA-24-METHYLCHOLESTA-7,22E-3BETA-OL
	516	64	0	0 412	(24S)-24-ETHYLCHOLESTA-5,25-DIEN-3BETA-O
44	999	106	999	0 398	24-METHYLCHOLESTA-5,24(28)-DIEN-3BETA-OL
	547	58	999	0 398	24-METHYLCHOLESTA-5,22E-DIEN-3BETA-OL
	521	61	999	0 426	GORGOST-5-EN-3BETA-OL
	527	39	999	0 426	(24Z)-24-PROPYLIDENECHOLESTA-5-EN-3BETA-
48	999	132	999	0 400	24-METHYLCHOLEST-5-EN-3BETA-OL
49	999	86	999	0 400	5ALPHA-24-METHYLCHOLEST-7-EN-3BETA-OL
	666	56	999	0 414	5ALPHA-24-ETHYLCHOLEST-7-EN-3BETA-OL
	529	45	999	0 386	5ALPHA-CHOLEST-7-EN-3BETA-OL
54	999	107	999	0 402	5ALPHA-24-METHYLCHOLESTAN-3BETA-OL
	584	59	999	0 416	5ALPHA-24-ETHYLCHOLESTAN-3BETA-OL
60	999	115	999	0 412	24-ETHYLCHOLESTA-5,22E-DIEN-3BETA-OL
	552	68	0	0 412	(25S)-24,27-DIMETHYLCHOLESTA-5,24(28)-DI
	547	46	999	0 412	23,24-DIMETHYLCHOLESTA-5,22-DIEN-3BETA-O
	522	58	999	0 412	5ALPHA-24-ETHYLCHOLESTA-7,22E-DIEN-3BETA
61	999	111	999	0 412	5ALPHA-24-ETHYLCHOLESTA-7,22E-DIEN-3BETA
	529	62	0	0 412	23-NOR-GORGOST-5-EN-3BETA-OL
	513	59	999	0 412	24-ETHYLCHOLESTA-5,22E-DIEN-3BETA-OL
	523	44	999	0 412	23,24-DIMETHYLCHOLESTA-5,22-DIEN-3BETA-O
63	999	86	0	0 410	24-ETHYLCHOLESTA-5,7,22E-TRIEN-3BETA-OL
	519	54	999	0 396	24-METHYLCHOLESTA-5,7,22E-TRIEN-3BETA-OL
68	999	123	999	0 412	24-ETHYLCHOLESTA-5,24(28)E-DIEN-3BETA-OL
	705	67	0	0 412	24-ETHYLCHOLESTA-5,24(28)Z-DIEN-3BETA-OL
	648	48	999	0 426	(24Z)-24-PROPYLIDENECHOLESTA-5-EN-3BETA-
	588	73	0	0 412	(24S)-24-ETHYLCHOLESTA-5,25-DIEN-3BETA-O
	581	68	0	0 412	23-NOR-GORGOST-5-EN-3BETA-OL
71	999	95	0	0 412	24-ETHYLCHOLESTA-5,24(28)Z-DIEN-3BETA-OL
	626	77	999	0 412	24-ETHYLCHOLESTA-5,24(28)E-DIEN-3BETA-OL
	648	48	999	0 426	(24Z)-24-PROPYLIDENECHOLESTA-5-EN-3BETA-
	581	68	0	0 412	23-NOR-GORGOST-5-EN-3BETA-OL
	512	63	0	0 412	(25S)-24,27-DIMETHYLCHOLESTA-5,24(28)-DI

Table V.

(cont.)

STEROL #	RANK			STEROLS MATCHED STEROL
74	999	124	0	0 412 (24S)-24-ETHYLCHOLESTA-5,25-DIEN-3BETA-O
	585	72	999	0 412 24-ETHYLCHOLESTA-5,24(28)E-DIEN-3BETA-OL
	536	66	0	0 412 (25S)-24,27-DIMETHYLCHOLESTA-5,24(28)-DI
	504	59	0	0 412 23-NOR-GORGOST-5-EN-3BETA-OL
75	999	84	999	0 412 23,24-DIMETHYLCHOLESTA-5,22-DIEN-3BETA-O
77	999	123	0	0 412 (25S)-24,27-DIMETHYLCHOLESTA-5,24(28)-DI
	598	70	0	0 412 23-NOR-GORGOST-5-EN-3BETA-OL
	573	66	999	0 412 24-ETHYLCHOLESTA-5,22E-DIEN-3BETA-OL
	560	69	999	0 412 24-ETHYLCHOLESTA-5,24(28)E-DIEN-3BETA-OL
	548	68	999	0 412 (24S)-24-ETHYLCHOLESTA-5,25-DIEN-3BETA-O
78	999	117	0	0 412 23-NOR-GORGOST-5-EN-3BETA-OL
	621	59	999	0 412 24-ETHYLCHOLESTA-5,24(28)Z-DIEN-3BETA-OL
	569	70	999	0 412 (25S)-24,27-DIMETHYLCHOLESTA-5,24(28)-DI
	560	69	999	0 412 24-ETHYLCHOLESTA-5,24(28)E-DIEN-3BETA-OL
	516	64	999	0 412 (24S)-24-ETHYLCHOLESTA-5,25-DIEN-3BETA-O
79	999	103	999	0 414 5ALPHA-23-NOR-GORGOSTAN-3BETA-OL
80	999	132	999	0 414 24-ETHYLCHOLEST-5-EN-3BETA-OL
	716	91	0	0 0 (24R,27S)-24,27-DIMETHYLCHOLEST-5-EN-3BE
82	999	101	999	0 416 5ALPHA-24-ETHYLCHOLESTAN-3BETA-OL
	523	56	999	0 402 5ALPHA-24-METHYLCHOLESTAN-3BETA-OL
86	999	105	999	0 416 5BETA-24-ETHYLCHOLESTAN-3BETA-OL (S01533-
88	999	84	999	0 414 5ALPHA-24-ETHYLCHOLEST-7-EN-3BETA-OL
	569	49	999	0 406 5ALPHA-24-METHYLCHOLEST-7-EN-3BETA-OL
89	999	101	999	0 412 24-ETHYLCHOLESTA-5,7-DIEN-3-BETA-OL
90	999	127	0	0 0 (24R,27S)-24,27-DIMETHYLCHOLEST-5-EN-3BE
	734	97	999	0 414 24-ETHYLCHOLEST-5-EN-3BETA-OL
	500	42	999	0 414 5ALPHA-24-ETHYLCHOLEST-7-EN-3BETA-OL
91	999	117	999	0 426 GORGOST-5-EN-3BETA-OL
	743	55	999	0 426 (24Z)-24-PROPYLIDENECHOLESTA-5-EN-3BETA-
	560	51	999	0 426 5ALPHA-GORGOST-7-EN-3BETA-OL
	516	61	999	0 426 24-ISOPROPYLCHOLESTA-5,22E-3BETA-OL (S01
	513	57	999	0 384 CHOLESTA-5,24-DIEN-3BETA-OL
92	999	91	999	0 426 5ALPHA-GORGOST-7-EN-3BETA-OL
93	999	123	999	0 428 5ALPHA-GORGOSTAN-3BETA-OL
94	999	74	999	0 426 (24Z)-24-PROPYLIDENECHOLESTA-5-EN-3BETA-
	538	63	999	0 426 GORGOST-5-EN-3BETA-OL
	505	48	999	0 412 24-ETHYLCHOLESTA-5,24(28)Z-DIEN-3BETA-OL
96	999	118	999	0 426 24-ISOPROPYLCHOLESTA-5,22E-3BETA-OL (S01
	529	62	999	0 426 GORGOST-5-EN-3BETA-OL

An important limitation of the file search system is then its inability to distinguish between variations in side chain alkylation. These various side chain alkylation patterns are very important with respect to biosynthetic processs. Since these sterols have different retention indices, this limitation has been overcome by searching a file of retention indices as well as mass spectra. A computer program for accurately calculating retention indices has been developed by William Yeager, Department of Genetics, Stanford University, and is applicable to the rapid analysis sequence. Michael Kohraman has prepared a file of carefully measured retention indices from samples used to compile the mass spectral file; (Table VI) therefore, the limitations concerning identification of isomeric side chain alkylation patterns have been reduced.

See p. 89a for Table VI.

TABLE VI

RETENTION INDICES OF STEROLS OF SP2250

MARINE STEROL NUCLEI

MARINE STEROL SIDE CHAINS

	2650 1							
	2892 2							
	3							
	2824 4							
	2972 5							
	2971 6		2927 7					
	3047 8							
	3282 9	3268 10	3317 11		3091 12			
	3258 13							
	3411 14							
	3319 15	3385 16	3325 17					
	3328 18	19	20	21	3219 22			
	3442 23	24						
	3361 25	3334 26	3383 27	3402 28	3245 29	3290 30	31	3322 32
		33						
	34	35	3364 36	37				
	3420 38	3497 39	3480 40	3492 41	3292 42			
	3468 43							
	3473 44	45	46		47			
	3445 48	3538 49	3474 50	51				
	52	53	54	55	3348 56	57	58	
	59							
	3492 60	3566 61		3557 63				
	3492 66	3624 67	3552 62		3376 64	65		
	3574 68	69	70					
	3601 71	3654 72	73					
	3555 74							
	3489 75				76			
	3568 77							
	3574 78		3560 79					
	3536 80	81	3560 82	83				
	3553 87	88		89	3427 84		85	86
	3578 90							
	3684 91	3755 92	3689 93					
	94							
	95							
	3487 96							
	3578 97							

Second, some sterols have very distinctive mass spectra with respect to the other spectra in the file, and no other spectrum ranks above 500 (for 17 spectra); however, the majority of spectra do show some similarities to other spectra in the file, i.e, have a cross rank > 500 with another sterol mass spectrum in the file. It is interesting that sterols which are saturated match only with other saturated sterols, sterols with one nuclear unsaturation match only with other sterols with one nuclear unsaturation, sterols with 2 nuclear unsaturations match only sterols with 2 nuclear unsaturations, and sterols with one nuclear and one side chain unsaturation (or ring junction) match only sterols possessing that property. The empirical ranking algorithm described previously has detected the number and general positions of unsaturation in the sterols. Therefore, if a new sterol is detected by the file search procedures then the general structural properties of the new sterol (number of nuclear and side chain double bonds) may be indicated by the structures of the sterols with which it is ranked even though the ranking values are very low.

The real utility of the search system will be in rapidly sorting a tremendous quantity of experimental data in an effort to reveal the sterols of novel structure. This is of tremendous utility because marine sterol mixtures are generally complex, containing over 40 sterols in some cases. However, once the sterol of novel structure is pointed out, then a careful analysis of the mass spectral fragmentation in terms of known processes must proceed. Rules generated via INTSUM, etc. analyses of the extensive marine sterol high resolution mass spectral files will help greatly by providing firm guidelines for the structural evaluations of the previously unencountered sterols.

4.2.2 Researchers Receiving Marine Sterol Data

Dr. J. B. Heather
The Upjohn Company
Chemical Process,
Rsch & Development
Kalamazoo, Mich.

Dr. Steven C. Welch
Dept of Chemistry
University of Houston
Houston, Texas 77004

Dr. Richard M. Wing
Univ of California
Riverside, Ca.

Prof. Paul J. Scheuer
University of Hawaii
2545 The Mall
Dept of Chemistry
Honolulu, Hawaii

Dr. Yuzura Shimizu
Univ of Rhode Island
College of Pharmacy
53 Fogarty
Kingston, R.I.

Dr. Maktoob Alam
University of Houston
College of Pharmacy
Dept. of Med. Chem.
and Pharmacognosy
Houston, Texas 77004

Dr. Ron Quinn
Roche Research Inst.
P. O. Box 255
Dee Why NSW 2099
AUSTRALIA

Dr. K. Ivanetich
Dept Physiol. & Med.
Biochemistry
Medical School
Observatory, Cape
SOUTH AFRICA

5 Carbon-13 Work

The work described in this section was accomplished in conjunction with work on structure elucidation and theory formation programs (sections 2 and 4). It is presented together here to make a more coherent presentation.

Carbon-13 nuclear magnetic resonance (CMR) has developed into an important tool for the structural chemist. A natural abundance CMR spectrum which is fully proton decoupled consists of a number of sharp peaks which correspond to the resonance frequencies in an applied magnetic field of the various types of carbon atoms present. A ^{13}C shift is the amount an observed peak is shifted from that of a reference peak, usually tetramethylsilane (TMS).

In last year's annual report we discussed an extension of Meta-DENDRAL which allowed the program to form rules in the domain of CMR spectroscopy. During the past year we continued work on this program, and wrote a second program which applies CMR rules to structure elucidation problems. Rules generated from a combined set of paraffins and acyclic amines have been used to successfully identify the ^{13}C NMR spectra of molecules not in the training set data. The introduction of a limited set of stereochemical terms to the rule generation procedure demonstrated the feasibility of extending the method to more complicated systems. A description of the rule formation and structure elucidation programs is given in [17]. Results are presented there for the combined set of paraffin and acyclic amines, as well as for a combined set of trans decalins and monohydroxylated androstanes.

5.1 Rule Formation Results

A set of rules was generated using a subset of the paraffin data from Lindeman and Adams [12] combined with a subset of the acyclic amine data from Eggert and Djerassi [13] Molecules with the empirical formula C_9H_{20} and $C_6H_{15}N$ were excluded from the training set for later use in testing the generality of the rules. The rule set was tested by generating all structural isomers with the empirical formulas C_9H_{20} (35 isomers) and $C_6H_{15}N$ (39 isomers), predicting the spectrum of each isomer, then ranking the predicted spectra by similarity to a known spectrum. The rank of the predicted spectra associated with the correct candidate structure provides an indication of the utility and

---

[12] Lindeman, L.P. and J.Q. Adams, Anal. Chem., (1971), 43,p. 1245.

[13] Eggert, H. and C. Djerassi, **J. Amer. Chem. Soc.** (1973),**95**,p. 3710.

validity of the generated rules. For the above test we used the 24 C_9H_{20} spectra available from the work of Lindeman and Adams. The predicted spectra of the 35 structural isomers were compared and ranked against each of these available spectra. The results of this ranking for C_9H_{20} as well as a similar test on $C_6H_{15}N$ are shown in Table VII.

Empirical Formula	Number of Candidates	Number of Spectra	Rank of Correct Structure (Freq of Correct Ranking) 1^{st}	2^{nd}	...6^{th}	..9^{th}
C_9H_{29}	35	24	20/24	3/24		1/24
$C_6H_{15}N$	39	11	8/11	2/11	1/11	

Table VII. Results of Structure Ranking

5.2 Adding Stereochemistry to the Rule Language

The work on the paraffins and acyclic amines requires only topological descriptors in the language of atom features. Because of the dependence of ^{13}C shifts on stereochemical features [14] [15] it is necessary to have the facility to include stereochemical terms when they are required. Substituents placed on systems which have static conformations such as trans decalin and androstane with trans ring fusions can be described in discrete terms. The terms we selected describe the orientation on the ring of the substituent as either axial or equatorial, and either alpha or beta. For instance, a substituent is beta in 10-methyl-trans-decalin if it is on the same side of the ring as the methyl group and alpha if on the opposite side of the ring from the methyl group. The rule generation program with the extension of the language to include these atom features was run on a combined set of trans decalins, 10-methyl-trans-decalols and monohydroxylated androstanes with trans ring fusions selected from the works of Grover and Stothers [16] and Eggert et. al. [17]

---

[14] Grover, S.H. and J.B. Stothers, Can. J. Chem. (1974),52,p. 870.

[15] Eggert, H., C. VanAntwerp, N. Bhacca, and C. Djerassi, J. Org. Chem.,(1976),41,p. 71.

[16] Grover, Op. cit.

[17] Eggert, Op. cit.

Sixty rules were generated to cover the 249 data peaks of 17 compounds. Samples of the rules generated are shown in Figure 11. The examination of these rules will show that they are useful for the chemist who wants to study contributions to the total shift as well as for structure elucidation.

See p. 94a for rules.

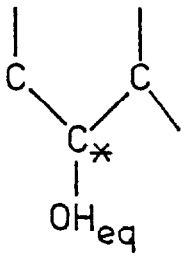
Figure 11. Sample rules constructed from decalins and hydroxy steroids with trans ring fusions. The '*' identifies the carbon atom to which the shift is assigned. is in ppm downfield from TMS.

5.3 Structure Elucidation

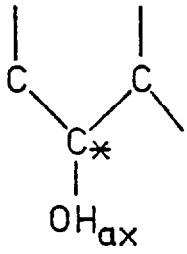
Molecular structure elucidation using CMR consists of using a set of rules which summarize the CMR behavior of a set of compounds to identify other unknown compounds within that or similar classes. The information which the chemist must supply to the structure elucidation program includes the empirical formula of the unknown as well as its observed spectrum. Two parameters may be set by the chemist to select the number of plausible structures to be determined, and to specify the error range in ppm which should be assigned to the rules to account for deficiencies in the training data, experimental error, solvent effects, etc. From this information and its store of CMR rules, the program assembles a set of structures which are plausible sources of the unknown spectrum.

Molecular structure elucidation is accomplished by our program by selecting a shift (peak) in the observed spectrum, then finding the rules which are possible explanations for this shift. The rules selected postulate partial substructures which

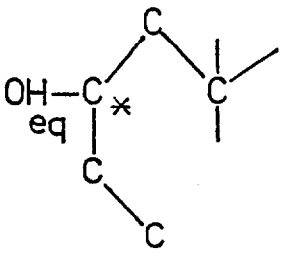
Alpha Carbon Rules



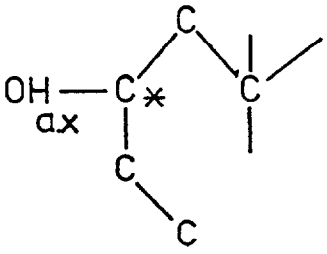
$\longrightarrow$ $70.0 \leq \delta(*) \leq 70.5$



$\longrightarrow$ $71.8 \leq \delta(*) \leq 72.5$

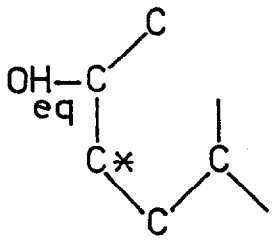


$\longrightarrow$ $66.9 \leq \delta(*) \leq 68.0$

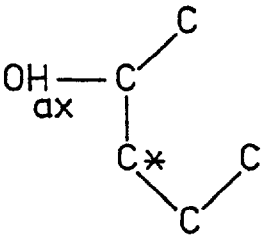


$\longrightarrow$ $67.6 \leq \delta(*) \leq 68.1$

Beta Carbon Rules

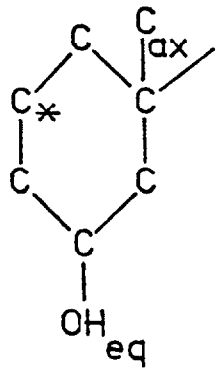


$\longrightarrow$ $35.6 \leq \delta(*) \leq 36.4$

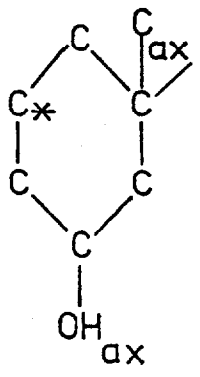


$\longrightarrow$ $33.9 \leq \delta(*) \leq 34.1$

Gamma Carbon Rules



$\longrightarrow$ $20.4 \leq \delta(*) \leq 20.5$



$\longrightarrow$ $16.9 \leq \delta(*) \leq 17.1$

might be in the molecule. These substructures are then assembled jigsaw puzzle fashion to construct the final molecule. Constraints stemming from both the observed spectrum and information associated with each rule are used to constrain the process so that only "reasonable" structures will be considered.

The structure elucidation program has been run on several test cases using unknown paraffin and acyclic amine spectra with reasonable success. This program is described in detail in [17].

5.4 Geometric Distortions in Steroids

For a given molecule, deviations between its observed ^{13}C NMR spectrum and its spectrum predicted from a set of empirical ^{13}C NMR rules is often explained in terms of geometric distortions. In order to examine the effect of geometric distortions on ^{13}C NMR shifts, Allinger's[18,19] molecular force field program has been used to model geometric distortions in mono-hydroxy- 5 alpha, 14-alpha androstanes. The net effect of many types of slight geometric distortions on the ^{13}C shift were examined in terms of the non-bonded interactions. The delta(alpha) and delta(beta) effects could be characterized in a few terms suggested by the non-bonded interactions. The results of this study were published in [16].

6 DATA COLLECTION AND DATA REDUCTION

6.1 DENDRAL GC/MS and MS Work

The following is a summary of the activities in the GC/MS lab for the past year. This work involves both development of the GC/MS Computer systems for both high and low resolution GC/MS applications and application of the existing system to mass spectral analyses of compounds of biomedical importance.

A) Low resolution GC/MS: (manual mode)

93 sterol mixtures (marine sterol extractions) for Dr. Djerassi's group. Identification of free sterols.

B) High Resolution GC/MS

---

[18] N.L. Allinger, M.T. Tribble, M.A. Miller and D.H. Wertz, J. Amer. Chem. Soc., 93, 1637 (1971).

[19] D.H. Wertz and N.L. Allinger, Tetrahedron, 30, 1579 (1974).

Total sample mixtures: 86

for:	1)	Dr. Djerassi	52
	2)	Genetics	25
	3)	Prof. Adlercreutz, Finland	9

1) Dr. Djerassi: all marine sterols, especially for library purposes and thesis of Bob Carlson.

2) Genetics: Urine extractions, channel-black and carbon-black, all for assistance in identification of unknown compounds whose structures could not be elucidated by low resolution mass spectral data coupled with library search.

3) Prof. Adlercreutz, Clinical Chemistry, University of Helsinki, Finland needed quantitation of a corticosteroid. Tests were made to find sensitivity limit with Aldosterone-TMS. 5 ug Alderstone-TMS was limitation. An unknown corticosteroid with a M+504 (low resolution spectrum) could not be identified by H. R. GC/MS due to amount of sample availability plus lack of sensitivity on our instrument. The sample was a substance occurring in patients who have no aldosterone, but still may have hypertension or hypokallemia.

High resolution MS

43 samples total:

for:	1)	Dr. Djerassi	29
	2)	Prof. Fringuelli, Italia	8
	3)	Prof. Nakano, Venezuela	6

1) Dr. Djerassi: Structure identification of new sterols plus terpenes.

2) Prof. Fringuelli, Perugia University, Perugia, Italia. Had 8 samples of furan, thiophen, selenophen and tellurophen derivatives for mass fragmentation studies. H. R. resolved all isotopes of each substance (up to 8 isotopes) and gave clear identification pattern. He is preparing and sending us more sets of compounds.

3) Prof. Nakano, Instituto Venezolano, Caracas, Venezuela, needed high resolution spectra of Oxadiazole derivatives for fragmentation studies, and successful identification of all six samples were possible.

Computerized MS (Incl. trials)

H. R. (R-10000) + GC/MS H.R. R-5000

Start Jan. 77.		Total
SO 1696 to 1921	DOS(*)	225
SO 1839 to 1859	DOS dublication nos.	20

SO 1960 to 2379	RT-11	419
SO 1883 to 1955	DOS	72
SO 2437 to 2477	RT-11	40
SO 1956 to 2031	DOS	75
SO 2479 to 2481	RT-11	2
SO 2032 to 2037	DOS	5
SO 2516 to 2907	RT-11	391
Total samples tested		1249

* DOS and RT-11 refer to the two different operating systems for the PDP-11 computer system. During the past year we have had to convert operating systems.

6.2 Collaborators Receiving the CLEANUP and HISLIB Programs

Following is an alphabetical list of people who have requested copies of the program for extracting better resolved mass spectra from GC/MS data, described in [10].

Dr. Craig Anderson
Gulf South Research Institute
P.O. Box 26500
New Orleans, Louisiana 70186

Dr. John B. Bagger
Department of Chemistry
Colorado State University
Fort Collins, Colorado 80521

Dr. Rod Britten
Jet Propulsion Laboratories
4800 Oak Grove Drive, 168-227
Pasadena, California 91103

Dr. Robert D. Brown
Bristol Laboratories
P. O. Box 657
Syracuse, New York 13201

Dr. Peter Bruck
Magyar Tudomanyos Akademia
Kozponti Kemiai Kutato Intezete
1088 Budapest Puskin u. 11-13.
Hungary

Dr. Lawrence Burkhard
Water Chemistry Laboratory
University of Wisconsin
660 North Park Street
Madison, Wisconsin 53706

Dr. Richard M. Caprioli
Dr. William E. Seifert, Jr.
Program in Biomolecular Analysis
Univ of Texas Medical School
P. O. Box 20708
Houston, Texas 77025

Dr. Henry E. Dayringer
Mail Zone V1A
Monsanto Agricultural Research
800 North Lindbergh Boulevard
St. Louis, Missouri 63166

Dr. James F. Elder
574 Building
Analytical Laboratories
Dow Chemical U.S.A.
Midland, Michigan 48640

Dr. W.K. Elkin
Department of Toxicology
Swedish Medical Research Council
Karolinska Institutet
S-104 01 Stockholm, Sweden

Dr. Paul V. Fennessey
B.F. Stolinsky Rsch Laboratories
Department of Pediatrics
Univ of Colorado Medical Ctr
4200 East Ninth Avenue
Denver, Colorado 80220

Dr. Claude Finn
School of Pharmacy
U. C. San Francisco Medical Ctr
San Francisco, California 94143

Dr. R. Fluckiger
Balzers Aktiengesellschaft fur
Hochvakuumtechnik und Dunne
Schichten
FL-9496 Balzers
Furstentum Liechtenstein

Dr. A.N. Freedman
Central Electricity
Research Laboratories
Kelvin Avenue, Leatherhead
Surrey, England

Dr. Nelson M. Frew
Chemistry Department
Woods Hole Oceanographic
Institution
Woods Hole, Massachusetts 02543

Dr. Richard Gans
Chemical Research Division
Bound Brook Laboratories
American Cyanamid Company
Bound Brook, New Jersey 08805

Mrs. E.M. Gomm
Department of Chemistry
University of Natal
P.O. Box 375, Pietermaritzburg
Natal, South Africa

Dr. Sydney M. Gordon
Chemistry Division
Atomic Energy Board
Private Bag 256, Pretoria
Republic of South Africa

Dr. Richard A. Graham
FSL
U. S. Army Natick Laboratories
Natick, Massachusetts 01760

Dr. Donald A. Griffin
Dept of Agricultural Chemistry
Oregon State University
Corvallis, Oregon 97331

Dr. William Haddon
Western Regional Research Center
U.S. Department of Agriculture
800 Buchanan Street
Albany, California 94710

Dr. P.T. Holland
Ministry of Agriculture
and Fisheries
Private Bag, Hamilton
New Zealand

Dr. I. Howe
Shell Biosciences Laboratory
Sittingbourne Research Centre
Sittingbourne, Kent ME9 8AG,
England

Akio Ide, Ph.D.
Ehime University
Agricultural Chemistry Dept
Matsuyama 790, Japan

Dr. J. B. Justice
Emory University
Atlanta, Georgia 30322

Dr. Graham S. King
Department of Chemical Pathology
Queen Charlotte's Hospital
Goldhawk Road
London, ENGLAND W6 OXG

Dr. Daniel R. Knapp
Department of Pharmacology
Medical Univ of South Carolina
80 Barre Street
Charleston,South Carolina 29401

Dr. H. Knoeppel
EURATOM - CCR
Casella Postale No. 1
Ispra, Italy

Dr. G. Knowles
Water Research Centre
Stevenage Laboratory
Elder Way, Stevenage
Hertfordshire SG1 1TH, England

Dr. Thomas Knudsen
Northrop Services
Box 12313
Research Triangle Park, N.C.

Dr. Douglas W. Kuehl
Mass Spectrometry Laboratory
Environmental Rsch Laboratory
6201 Congdon Boulevard
Duluth, Minnesota 55804

Dr. Ake Lundin
LKB-PRODUKTER AB
Molecular Analysis Division
S-161 25 Bromma 1
Sweden

Dr. John L. MacDonald
Central Research
Ralston Purina Company
Checkerboard Square
St. Louis, Missouri 63188

Dr. R.G.A.R. Maclagan
Department of Chemistry
University of Canterbury
Christchurch 1, New Zealand

Dr. John C. Marshall
Department of Chemistry
The University of North Carolina
Chapel Hill, N.C.

Dr. R. A. F. Matheson
Chemistry Section
Environmental Protection Service
5151 George Street
Halifax, Nova Scotia CANADA

Dr. James A. McCloskey, Jr.
Professor, Biomedical Chemistry
Dept. Biopharmaceutical Sciences
University of Utah
Salt Lake City, Utah 84112

Dr. Ingolf Meineke
Fachbereich Chemie
Philipps Universitaet
3550 Marburg/Lahn, Lahnberge
WEST GERMANY

Dr. Roy O. Morris
Dept. Agricultural Chemistry
Oregon State University
Corvallis, Oregon 97331

Dr. James E. Oberholtzer
Arthur D. Little, Inc.
Acorn Park
Cambridge, Massachusetts 02140

Mr. Andrew Pallos
Aerospace Corporation
P.O. Box 92957
Los Angeles, California 90009

Mr. Dan Pearce
Orange Co Sheriff-Coroner Dept
550 N. Flower Street
Santa Ana, California 92702

Dr. William R. Penrose
Newfoundland Biological Station
3 Water St. East
St. John's, Newfoundland
AlC 1A1

Dr. Ronald D. Plattner
Northern Regional Research Lab.
U.S. Department of Agriculture
Peoria, Illinois 61604

Ken Pocek
Scientific Instruments Division
Hewlett-Packard Company
1601 California Avenue
Palo Alto, California 94304

Dr. Philip W. Ryan
Battelle Pacific Northwest
Laboratories, 329 Bldg.
Battelle Boulevard
Richland, Washington 99352

Dr. Robert S. Schroeder
Gulf Oil Chemicals Company
P. O. Box 2900
Shawnee Mission, Kansas 66201

Dr. J. Scrivens
Imperial Chemical Industries
PO Box 90 Wilton
Middlesbrough Cleveland
TS6 8JE England

Dr. Walter M. Shackelford
Analytical Chemistry Branch
Environmental Rsch Laboratory
Athens, Georgia 30601

Dr. M.A. Shaw
Unilever Research
Port Sunlight Laboratory
Port Sunlight
Wirral, Merseyside L62 4XN,
England

Dr. Jacob Shen
The Standard Oil Company
4440 Warrensville Center Road
Cleveland, Ohio 44128

Dr. M. M. Siegel
FMC Corporation
Chemical Group
Box 8
Princeton, New Jersey 08540

Dr. G. P. Slater
National Rsch Council of Canada
Prairie Regional Laboratory
110 Gymnasium Road,
University Campus
Saskatoon, Saskatchewan CANADA

Dr. Carroll A. Smith
Div of Chemical Oceanography
University of Miami
4600 Rickenbacker Causeway
Miami, Florida 33149

Dr. H. J. Stoklosa
Central Rsch & Development Dept
E. I. du Pont de Nemours & Co.
Wilmington, Delaware 19898

Dr. F. Street
AEI Scientific Apparatus, Ltd.
Barton Dock Road
Urmston, Manchester M31 2LD
England

Dr. Robert M. Supnik
Massachusetts Computer Assoc.
26 Princess Street
Wakefield, Massachusetts 01880

Dr. H. G. J. Teisman
Netherlands Institute for
Dairy Research
Post Office Box 20
EDE, NETHERLANDS

Dr. Gareth Templeman
Rsch and Development Lab.
The Pillsbury Company
311 Second Street Southeast
Minneapolis, Minnesota 55414

Dr. Ernst Weber
Varian MAT
P.O. Box 144062
Bremen, West Germany

Dr. J. Wyatt
Code 6110
Naval Research Laboratory
Washington, D. C. 20375

7 APPENDICES

7.1 Appendix A
The Structures of All 4-Dimethyl Marine Sterols Reported to the Beginning of 1977.

Each sterol is given a unique number which is used in subsequent discussions in the text.

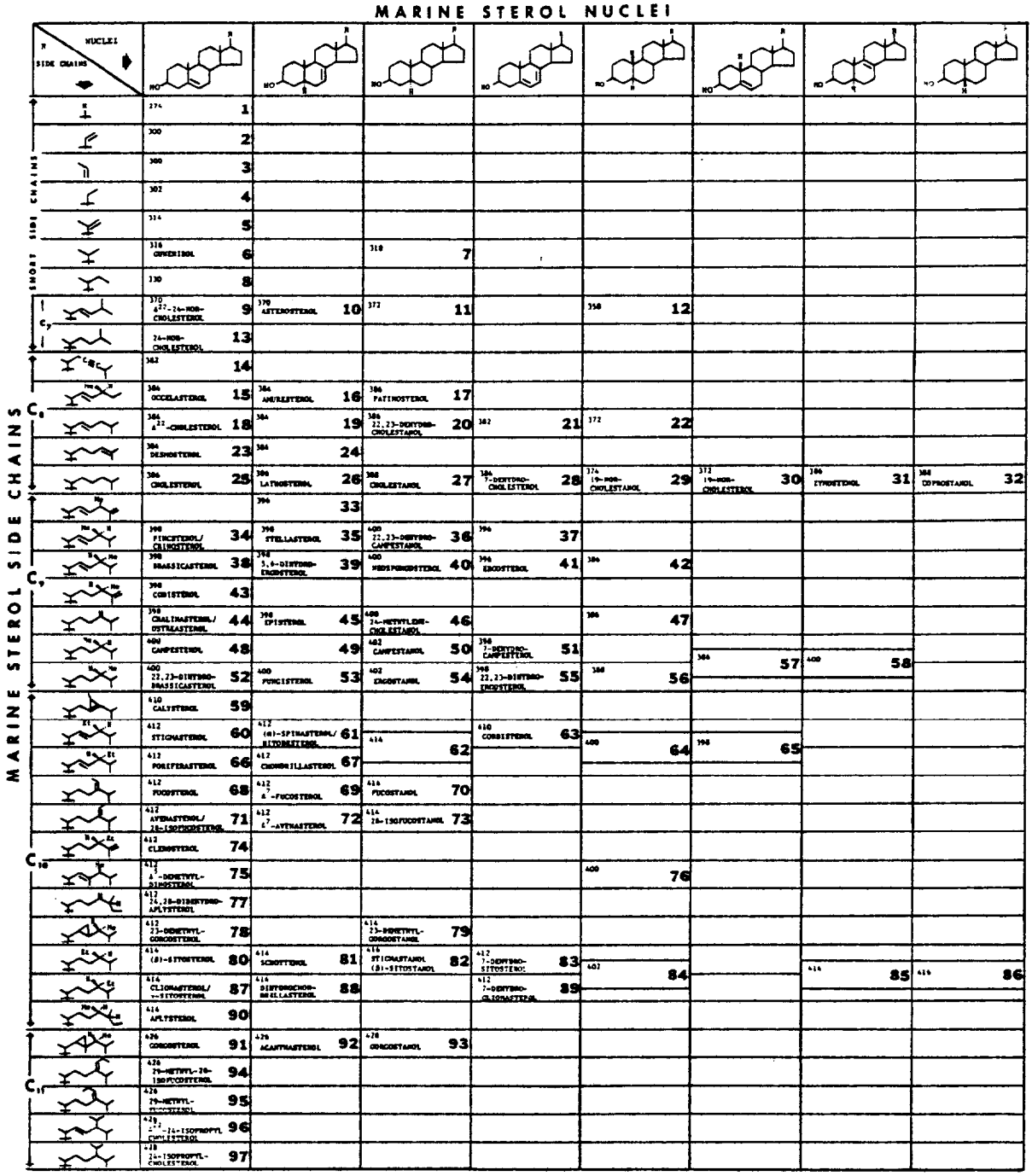
The molecular weight (M^+ and common trivial name are given for each sterol.

The nuclei all possess alternating trans-anti stereochemistry at the ring junctures, except the 5* stanols (farthest right hand column) which possess a cis-A,B ring fusion.

The number of carbon atoms in the side chains is indicated along the left hand border.

See p. 102a for Appendix I.

MARINE STEROL NUCLEI



7.2 Appendix B
Sources of Sterol Mass Spectra.

Sterol : names listed on the following two pages indicate spectra that were obtained from the old mass spectral files of Prof. Carl Djerassi. The spectra are divided into two groups: (1) sterols that are known to occur in marine sources, i.e. "4-demethyl marine sterol mass spectra" and (2) all other sterol mass spectra from those files grouped under "4-demethyl synthetic sterol mass spectra". The original CD number is given. These spectra were incorporated in the National Institutes of Health MSSS mass spectral data bank, which is available internationlly to researchers employing mass spectral identifications systems. See: S. R. Heller, Biomed. Mass., 1, 207 (1974).

All other mass spectra listed in Appendix 1 have been obtained by running mass spectra of authentic samples provided by researchers from around the world. The samples were either pure compounds or mixtures requiring subsequent purifications here. I would,therefore, like to join Professor Djerassi in thanking these researchers.

Dr. Aringer
(Karolinska Siukhuset, Stockholm)

Dr. J. T. Baker and Dr. R. J. Wells
(Roche Research Institute of Pharmacology, Australia)

Dr. M. Barbier
(Institut de Chimie des Substances Naturelles, France)

Dr. L. J. Goad
(University of Liverpool, England)

Dr. A. Kanazawa
(University of Kagoshima, Japan)

Dr. B. A. Knights
(University of Glasgow, Scotland)

Dr. M. Kobayashi
(Hokkaido University, Japan)

Dr. J. Mathieu
(Roussel-UCLAF Research Laboratories)

Dr. P. J. Scheuer
(University of Hawaii)

Dr. F. J. Schmitz
(University of Oklahoma)

Dr. R. H. Thomson
(University of Aberdeen)

Dr. A. J. Weinheimer
(University of Oklahoma)

I would also like to thank Willian A. Dow, Stanford University, for samples of aplysterol 90 and didehydroaplysterol 77 which he isolated from Verongia fistularis.

4-DEMETHYL SYNTHETIC STEROL MASS SPECTRA FROM THE FILES OF
CARL DJERASSI JAN. 7, 1976
(ALL SPECTRA RECORDED AT 70EV)

	SD#	CAT#	CD#	MW	MS	FORMULA	STEROL
1		401	05203	276	CEC-103	C19H320	5ALPHA-ANDROSTAN-3BETA-OL
2		→ 99	16309	300	AEI-MS9	C21H320	(17(20)Z)-PREGNA-5,17(20)-DIEN-3BETA-OL
3		→221	99999	302	MAT-CH4	C21H340	PREG-5-EN-3BETA-OL
4		219	09576	316	AEI-MS9	C22H340	23,24-DINOR-CHOL-20-EN-3BETA-OL
5		---	11111	328	U	C23H340	24-NOR-CHOLA-5,22-DIEN-3BETA-OL
6		5673	20636	342	U	C24H380	(22E)-CHOLA-5,22-DIEN-3BETA-OL
7		5672	20637	342	U	C24H380	(22Z)-CHOLA-5,22-DIEN-3BETA-OL
8		78	18697	346	AEI-MS9	C24H420	5BETA-CHOLAN-3BETA-OL
9		76	18713	346	U	C24H420	5BETA-CHOLAN-3ALPHA-OL
10		5671	20641	356	U	C25H400	(22Z)-26,27-DINOR-CHOLESTA-5,22-DIEN-3BETA-OL
11		5670	20639	356	U	C25H400	(22E)-26,27-DINOR-CHOLESTA-5,22-DIEN-3BETA-OL
12		5669	20659	370	U	C26H420	(22Z)-24-NOR-CHOLESTA-5,22-DIEN-3BETA-OL
13		5345	99690	370	U	C26H420	24-NOR-CHOLESTA-5,25-DIEN-3BETA-OL
14		5344	99691	370	U	C26H420	24-NOR-CHOLESTA-5,23-DIEN-3BETA-OL
15		4691	18661	382	MAT-CH4	C27H420	CHOLESTA-5,20(21),24-TRIEN-3BETA-OL
16		4692	18659	382	MAT-CH4	C27H420	(17(20)E)-CHOLESTA-5,17(20),24-TRIEN-3BETA-OL
17		5667	20525	384	U	C27H440	(22Z)-27-NOR-24-METHYLCHOLESTA-5,22-DIEN-3BETA-OL
18		4693	18803	384	MAT-CH4	C27H440	(20(22)E)-CHOLESTA-5,20(22)-DIEN-3BETA-OL
19		4698	18723	384	U	C27H440	CHOLESTA-5,20(21)-DIEN-3BETA-OL
20		3299	99812	386	U	C27H460	CHOLEST-8(14)-EN-3BETA-OL
21		231	06032	398	CEC-103	C28H460	24-METHYLCHOLESTA-5,24(25)-DIEN-3BETA-OL
22		234	09180	400	MAT-CH4	C28H480	24-METHYLCHOLEST-8(14)-EN-3BETA-OL
23		3518	19479	410	U	C29H460	22,23-METHYLENE-24-METHYLCHOLESTA-5,24(28)-DIEN-3β-O
24		4785	19338	412	MAT-CH4	C29H480	(22E)-24-DIMETHYLCHOLESTA-5,22-DIEN-3BETA-OL

SB# = SAMPLE BOX NUMBER
CAT# = TABLET CATALOG NUMBER
CD# = CARL DJERASSI MASS SPECTRUM NUMBER
MW = MOLECULAR WEIGHT
MS = MASS SPECTROMETER

"→" indicates that spectrum has subsequently been moved to the marine file.

4-DEMETHYL MARINE STEROL MASS SPECTRA FROM THE FILES OF CARL DJERASSI JAN, 7, 1976

(ALL SPECTRA RECORDED AT 70EV)

	SB#	CAT#	CD #	MW	MS	FORMULA	STEROL
1	A-2	5668	20660	370	U	C26H420	24-NOR-CHOLESTA-5,22-DIEN-3BETA-OL
2	A-1	5339	18380	372	AEI-MS9	C26H440	24-NOR-CHOLEST-5-EN-3BETA-OL
3	E-2	4739	99754	372	U	C26H440	→(22E)-19-NOR-5ALPHA,10BETA-CHOLEST-22-EN-3BETA-OL
4	A-15	100	16746	384	AEI-MS9	C27H440	CHOLESTA-5,7-DIEN-3BETA-OL
5	A-8	237	05570	384	MAT-CH4	C27H440	CHOLESTA-5,24-DIEN-3BETA-OL
6	B-2	103	16793	384	AEI-MS9	C27H440	➧5ALPHA-CHOLESTA-7,9(11)-DIEN-3BETA-OL
7	A-5	4732	17657	386	U	C27H460	CHOLEST-5-EN-3BETA-OL
8	A-11	101	16778	386	AEI-MS9	C27H460	5ALPHA-CHOLEST-7-EN-3BETA-OL
9	A-18	2509	18633	388	U	C27H480	5ALPHA-CHOLESTAN-3BETA-OL
10	A-18	406	05557	388	CEC-103	C27H480	5ALPHA-CHOLESTAN-3BETA-OL
11	B-5	4748	20063	398	AEI-MS9	C28H460	(22E)-24-METHYLCHOLESTA-5,22-DIEN-3BETA-OL
12	B-5	2455	15214	398	U	C28H460	(22E)-24-METHYLCHOLESTA-5,22-DIEN-3BETA-OL
13	B-10	232	06060	398	MAT-CH4	C28H460	24-METHLYCHOLESTA-5,24(28)-DIEN-3BETA-OL
14	B-13	233	09125	398	MAT-CH4	C28H460	(22E)-24-METHYL-5ALPHA-CHOLESTA-7,22-DIEN-3BETA-OL
15	C-7	3503	15281	412	AEI-MS9	C29H480	(22E)-24-ETHYLCHOLESTA-5,22-DIEN-3BETA-OL
16	C-7	2454	14916	412	U	C29H480	(22E)-24-ETHYLCHOLESTA-5,22-DIEN-3BETA-OL
17	C-20	236	06052	412	MAT-CH4	C29H480	(22E)-24-ETHYL-5ALPHA-CHOLESTA-7,22-DIEN-3BETA-OL
18	C-9	2438	12489	412	U	C29H480	(24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL
19	C-9	3406	11257	412	AEI-MS9	C29H480	(24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL
20	C-9	4742	99752	412	U	C29H480	(24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL
21	C-9	238	06237	412	AEI-MS9	C29H480	(24E)-STIGMASTA-5,24(28)-DIEN-3β-ol
22	D-9	4738	18100	416	U	C29H520	24-ETHYL-5ALPHA-CHOLESTAN-3BETA-OL
23	D-14	2450	13915	426	AEI-MS9	C30H500	GORGOSTEROL
24	D-14	4733	17659	426	U	C30H500	GORGOSTEROL
25	D-12	4740	19975	426	U	C30H500	(24Z)-24-PROPYLIDENECHOLEST-5-EN-3BETA-OL

SB# = SAMPLE BOX NUMBER
CAT# = TABLET CATALOG NUMBER
CD# = CARL DJERASSI MASS SPECTRUM NUMBER
MW = MOLECULAR WEIGHT
MS = MASS SPECTROMETER

➧ indicates spectrum has been moved to the synthetic file

"→" indicates spectrum has subsequently been deleted because of poor quality.

References

1. Bruce G. Buchanan and Dennis H. Smith, "Computer Assisted Chemical Reasoning," in E.V. Ludena, N.H. Sabelli and A.C. Wahl (eds.), Computers in Chemical Education and Research, New York: Plenum Press, 1977. P. 401

2. Bruce G. Buchanan, "Issues of Representation in Conveying the Scope and Limitations of Intelligent Assistant Programs," in D. Michie (ed.), Machine Intelligence 9, forthcoming.

3. Bruce G. Buchanan and Tom Mitchell. "Model-Directed Learning of Production Rules," in D.A. Waterman and F. Hayes-Roth (eds.), Pattern-Directed Inference Systems, New York: Academic Press, forthcoming.

4. Bruce G. Buchanan and Edward A. Feigenbaum, "DENDRAL and Meta-DENDRAL: Their Applications Dimension," Artificial Intelligence, forthcoming.

5. Raymond E. Carhart and Dennis H. Smith, "Applications of Artificial Intelligence for Chemical Inference XX. Intelligent Use of Constraints in Computer-Assisted Structure Elucidation", Computers and Chemistry, 1, 79 (1976).

6. Raymond E. Carhart, "A Model-Based Approach to the Teletype Printing of Chemical Structures," Journal of Chemical Information and Computer Sciences, 16, 82, 1976.

7. R.E. Carhart, T.H. Varkony, and D.H. Smith, "Computer Assistance for the Structural Chemist," in "Computer-Assisted Structure Elucidation," D.H. Smith, (ed.), American Chemical Society, Washington, D.C., 1977, p. 126.

8. C.J. Cheer, D.H. Smith and C. Djerassi, and B. Tursch, J.C. Braekman and D. Daloze, "Applications of Artificial Intelligence for Chemical Inference XXI: The Computer-Assisted Identification of [+]Palustrol in the

Marine Organism Cespitularia sp., aff. Subviridis", Tetrahedron, 32, 1807 (1976).

9. C. Djerassi, R. M. K. Carlson, S. Popov and T. H. Varkony. Sterols from Marine Sources. In press.

10. R. G. Dromey, Mark J. Stefik, Thomas C. Rindfleisch, and Alan M. Duffield, "Extraction of Mass Spectra Free of Background and Neighboring Component Contributions from Gas Chromatography/Mass Spectrometry Data," Analytical Chemistry, 48, 1368, August 1976.

11. T.M. Mitchell and G.M. Schwenzer, "Applications of Artificial Intelligence for Chemical Inference XXV. A Computer Program for Automated Empirical 13C NMR Rule Formation," Organic Magnetic Resonance, forthcoming.

12. Tom M. Mitchell, "Version Spaces: A Candidate Elimination Approach To Rule Learning," Proceedings of the Fifth IJCAI, 1, 305, August 1977.

13. James G. Nourse, "Generalized Stereoisomerization Modes," Journal of the American Chemical Society, 99, 2063, 1977.

14. S. Popov, R. M. K. Carlson, A-M. Wegmann and C. Djerassi. Occurrence of 19-Nor Cholesterol and Homologs in Marine Animals. Tetrahedron Lett., 3491 (1976).

15. S. Popov, R. M. K. Carlson, A-M. Wegmann and C. Djerassi. Minor and Trace Sterols in Marine Invertebrates. 1. General Methods of Analysis. Steroids, 28, 699 (1976).

16. Gretchen M. Schwenzer, "Applications of Artificial Intelligence for Chemical Inference. XXVI Analysis of C-13 NMR for Mono-Hydroxy Steroids Incorporating Geometric Distortions," Journal of Organic Chemistry, forthcoming.

17. Gretchen M. Schwenzer and Tom M. Mitchell, "Computer Assisted Structure Elucidation Using Automatically Acquired 13C NMR Rules," in D. Smith, (ed.), Computer Assisted Structure Elucidation, ACS Symposium Series, Vol. 54:58, 1977.

18. D.H. Smith, (ed.), American "Computer-Assisted Structure Elucidation," Chemical Society, Washington, D.C., 1977.

19. D.H. Smith, M. Achenbach, W.J. Yeager, P.J. Anderson, W.L. Fitch, and T.C. Rindfleisch, "Quantitative Comparison of Combined Gas Chromatographic/Mass Spectrometric Profiles of Complex Mixtures," Anal. Chem., 49, 1623 (1977).

20. D.H. Smith and P.C. Jurs, "Prediction of 13C NMR Chemical Shifts," J. Am. Chem. Soc., submitted for publication.

21. Dennis H. Smith and Raymond E. Carhart, "Structure Elucidation Based on Computer Analysis of High and Low Resolution Mass Spectral Data," in M.L. Gross (ed.), Proceedings of the Symposium on Chemical Applications of High Performance Spectrometry, Washington, D.C.: American Chemical Society, in press.

22. T.H. Varkony, R.E. Carhart, and D.H. Smith, "Applications of Artificial Intelligence for Chemical Inference XXIII. Computer-Assisted Structure Elucidation. Modelling Chemical Reaction Sequences Used in Molecular Structure Problems," in W.T. Wipke, (ed.), Computer-Assisted Organic Synthesis, Washington, D.C.: American Chemical Society, 1977.

23. Tomas H. Varkony, Raymond E. Carhart, and Dennis H. Smith, "Computer Assisted Structure Elucidation, Ranking of Candidate Structures, Based on Comparison Between Predicted and Observed Mass Spectra," in Proceedings of the Twenty-Fifth Annual Conference on Mass Spectrometry and Allied Topics, Washington, D.C., 1977.

24. Tomas Varkony, Dennis Smith, and Carl Djerassi, "Computer-Assisted Structure Manipulation: Studies in the Biosynthesis of Natural Products," Tetrahedron, forthcoming.

25. T.H. Varkony, R.E. Carhart, D.H. Smith, and C. Djerassi, "Computer-Assisted Simulation of Chemical Reaction Sequences. Applications to Problems of Structure Elucidation," J. Am. Chem. Soc., submitted for publication.

26. Annemarie Wegmann, "Variations in Mass Spectral

Fragmentation Produced by Active Sites in a Mass Spectrometer Source," Analytical Chemistry, forthcoming.

The undersigned agrees to accept responsibility for the scientific and technical conduct of the project and for provision of required progress reports if a grant is awarded as the result of this application.

1/26/78
Date

Principal Investigator or
Program Director